DEEP LEARNING IN MEDICAL IMAGE ANALYSIS: EFFICIENT USE OF DATA AND RADIOLOGICAL EXPERTISE

Doctoral Dissertation by
Satheshkumar Kaliyugarasan

Thesis submitted for
the degree of Philosophiae Doctor (PhD)
in
Computer Science:
Software Engineering, Sensor Networks and Engineering Computing

Department of Computer Science,
Electrical Engineering and Mathematical Sciences

Faculty of Engineering and Science

Western Norway University of Applied Sciences

April, 2023

Series of dissertation submitted to
the Faculty of Engineering and Science,
Western Norway University of Applied Sciences.

ISBN: 978-82-8461-023-8

Author: Satheshkumar Kaliyugarasan
Title: Deep learning in medical image analysis: efficient use of data and radiological expertise

Printed production:
Molvik Grafisk / Western Norway University of Applied Sciences

Bergen, Norway, 2023

PREFACE

The author of this thesis has worked as a Ph.D. research fellow in the Data Science research group at the Department of Computer Science, Electrical Engineering, and Mathematical Sciences at Western Norway University of Applied Sciences (HVL). The position was funded by the Western Norway Regional Health Authority (Helse Vest RHF), project F-12532. He has been enrolled in the Ph.D. program in Computer Science: Software Engineering, Sensor Networks, and Engineering Computing.

The research presented in this thesis has been conducted at Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway. Part of the research has been done in collaboration with the Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Sweden.

This thesis is organized into two parts. Part I provides an overview of the relevant field and the background for the articles in the thesis, including a summary of the works. Part II consists of a collection of published and peer-reviewed research papers and a manuscript.

Paper A Kaliyugarasan, Satheshkumar and Lundervold, Alexander Selvikvåg. fastMONAI: a low-code deep learning library for medical image analysis. Manuscript, April 2023

Paper B Kaliyugarasan, Satheshkumar, Kociński, Marek, Lundervold, Arvid and Lundervold, Alexander Selvikvåg. 2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using fastai. In Proceedings of the of the 33rd Norwegian Informatics Conference (NIK), 23 November 2020

Paper C Kaliyugarasan, Satheshkumar, Lundervold, Arvid and Lundervold, Alexander Selvikvåg. Pulmonary nodule classification in lung cancer from 3D thoracic CT scans using fastai and MONAI. In International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), Volume 6, Number 7, 4 May 2021.

Paper D Hodneland, Erlend, Kaliyugarasan, Satheshkumar, Wagner-Larsen, Kari Strøno, Lura, Njål, Andersen, Erling, Bartsch, Hauke, Smit, Noeska, Halle, Mari Kyllesø, Krakstad, Camilla, Lundervold, Alexander Selvikvåg and Haldorsen, Ingfrid Salvesen. Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer. In Cancers, Volume 14, Number 10, 11 May, 2022.

Paper E Kaliyugarasan, Satheshkumar, Dagestad, Magnhild H., Papalini, Evin I., Andersen, Erling, Zwart, John-Anker, Brisby, Helena, Hebelka, Hanna, Ansgar, Espeland, Lagerstrand, Kerstin M. and Lundervold, Alexander Selvikvåg. Multi-Center CNN-based spine segmentation from T2w MRI using small amounts of data. To appear in the Proceedings of the of the 20th IEEE International Symposium on Biomedical Imaging (ISBI), 18-21 April 2023.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Assoc Prof. Alexander Selvikvåg Lundervold. Without your guidance and constant feedback this thesis would not have been achievable. Our discussions were often filled with excitement, occasionally leading us to get carried away. I truly appreciate the opportunity to have worked under your mentorship and look forward to continuing our collaboration in the future.

I would also like to extend my thanks to my co-supervisors, Prof. Ingfrid S. Haldorsen and Assoc. Prof. Erlend Hodneland.

Ingfrid, your insightful suggestions, and encouragement have been invaluable, and I look forward to continuing as a researcher in your group, primarily focusing on gynecological cancer.

Erlend, I'm grateful for the collaboration we shared during the cervical cancer project, and for your valuable input during the finalization of my thesis.

To Assoc. Prof. Hauke Bartsch and Zhanbolat Satybaldinov, I look forward to continuing our work on implementing deep learning applications into the hospital infrastructure.

I would also like to express my gratitude to Prof. Emeritus Arvid Lundervold, for fruitful and interesting discussions.

Thanks to all my colleagues at MMIV for creating such a joyful professional work environment. A special thanks to Njål Lura and Heidi Espedal (even though you moved to Australia), for the camaraderie, coffee chats, and wealth of wisdom you both shared.

To my friends, thank you for your understanding and moral support throughout this challenging yet rewarding journey. A special mention goes to Fredrik Mørk for the epic bike rides we enjoyed together during the pandemic.

Finally, I would like to express my heartfelt gratitude to my family, especially my parents, Kaliyugarasan Thirunavukkarasu and Jayanthi Kaliyugarasan, for their unconditional love and encouragement. I am also grateful to my siblings, Thinesh Kaliyugarasan and Jeelaja Kaliyugarasan, for their continuous support and motivation.

ABSTRACT

Deep learning (DL), a branch of artificial intelligence (AI), has experienced significant growth and advancements over the past decade and has shown great potential in various sectors, including the medical domain. The goal that drives deep learning research for medical applications is the development of tools that can enhance the accuracy and efficiency of diagnosis, reduce medical costs, and streamline and improve diagnostic processes through a greater degree of precision medicine, with better prognostics and stratification of therapy.

In modern medicine, radiology has become increasingly important, with medical imaging playing a critical role in detecting, diagnosing, and treating various diseases. Simultaneously, there is a shortage of qualified medical specialists, i.e., radiologists.

The potential of deep learning for medical image analysis is evident; however, much of the excitement around the applications is rooted in retrospective studies. In practice, only a limited number of deep learning-based studies have progressed to deployment in clinical care. Moreover, at least part of the field seems to be facing a reproducibility crisis.

The reasons for this are multiple, including technical debt, overfitting models, selection bias, and heavy preprocessing of data sets in the scientific community, not properly reflecting clinical diversity and local variations. These issues can be attributed, in part, to the insufficient collaboration between the medical and data science communities. To overcome these obstacles and fully realize the benefits of data-driven medical imaging, it is crucial to foster interdisciplinary collaboration.

As one possible remedy, deep learning frameworks tailored to medical imaging can help foster interdisciplinary collaboration, facilitate rapid iterative development, and support reproducible research. Such frameworks can make it easier for domain experts to join in on method development and for other researchers to verify the validity of the reported results and build upon existing work. This can help accelerate the integration of deep learning-based solutions into clinical practice.

To address these challenges and promote the integration of cutting-edge deep learning-based solutions into clinical practice, Medical Open Network for Artificial Intelligence (MONAI) provides an open-source PyTorch-based deep learning framework to support medical data, with a particular focus on imaging applications. Following best practices for software development, MONAI provides an easy-to-use, well-documented, and well-tested software framework freely available to all interested researchers via `https://monai.io/`.

In this thesis, we present `fastMONAI`, a low-code Python-based open-source deep learning library built on frameworks from MONAI. The library incorporates several best practices and state-of-the-art techniques by integrating capabilities from MONAI with two other powerful libraries: `fastai` and `TorchIO`, along with custom-made modules.

`fastMONAI` provides a high-level API that simplifies the process of data loading, preprocessing, training, and result interpretation, allowing researchers to spend less time on coding and focus more on the challenges within each project. Despite its

high-level interface, `fastMONAI` maintains the customization and flexibility of `fastai`, enabling experienced practitioners to incorporate custom extensions when needed.

The development and evaluation of `fastMONAI` have been conducted using both public and clinical study data involving multiple patient groups, radiological domains, and organ systems, including identifying the brain from surrounding tissue and structures (Paper B), lung cancer (Paper C), gynecological cancer (Paper D), and low back pain (Paper E). Each patient group requires accurate and efficient medical imaging analysis for diagnosis and treatment planning. Our results in this thesis demonstrate promising improvements in diagnostic accuracy and streamlined workflows.

However, to thoroughly evaluate the models, it is crucial to integrate them into real-world workflows and study their performance in realistic contexts. In this thesis, we found that the flexibility and the user-friendly API of `fastMONAI` facilitate the integration of trained models into clinical infrastructure (see Figure 4.5). This is explored further in ongoing and future work building on the thesis results.

SAMMENDRAG

Dyplæring (DL), en underkategori av kunstig intelligens (KI), har opplevd betydelig vekst og utvikling det siste tiåret. Dette har åpnet et stort potensial innen ulike sektorer, inkludert i helsevesenet. Målet med dyplærings-forskning for medisinske applikasjoner er å utvikle verktøy som kan forbedre nøyaktigheten og effektiviteten av diagnostisering, redusere medisinske kostnader og effektivisere og forbedre diagnostiske prosesser gjennom en større grad av presisjonsmedisin, med bedre prognostikk og stratifisering.

Radiologi har blitt stadig viktigere i moderne medisin, der medisinsk bildediagnostikk spiller en avgjørende rolle i deteksjon, diagnostisering og behandling av ulike sykdommer. Samtidig er det en mangel på kvalifiserte medisinske spesialister, i.e., radiologer.

At det er et potensial for dyplæring innen medisinsk bildeanalyse er åpenbart. Mye av entusiasmen rundt anvendelsene er imidlertid basert på retrospektive studier. Kun et fåtall studier basert på dyplæring har handlet om løsninger innlemmet i klinisk praksis og arbeidsflyt. I tillegg ser det ut til at deler av feltet står overfor en reproduserbarhetskrise.

Årsakene til dette kan være mange, blant annet teknisk gjeld (*technical debt*), overtilpassede modeller, seleksjonsskjevhet og omfattende preprossesering av datasett i det vitenskapelige miljøet som ikke gjenspeiler klinisk mangfold og lokale variasjoner. Problemene kan delvis skyldes utilstrekkelig samarbeid mellom medisinske og datavitenskapelige miljøer. For å takle disse utfordringene og fullt ut realisere fordelene med datadrevet medisinsk bildebehandling, er det viktig å fremme tverrfaglig samarbeid.

Ett mulig middel er å utvikle dyplærings-rammeverk for medisinsk bildediagnostikk som fremmer tverrfaglig samarbeid, legger til rette for rask iterativ utvikling og støtter reproduserbar forskning. Dette gjør det enklere for domeneeksperter å delta i metodeutvikling og for andre forskere å verifisere gyldigheten av de rapporterte resultatene og bygge videre på eksisterende arbeid. Slikt kan bidra til å akselerere integrasjonen av dyplæringsbaserte løsninger i klinisk praksis.

For å møte disse utfordringene og fremme integrasjonen av banebrytende dyplærings-baserte løsninger i klinisk praksis, tilbyr *Medical Open Network for Artificial Intelligence* (MONAI) et PyTorch-basert dyplærings-rammeverk med åpen kildekode rettet mot medisinske data, med et spesielt fokus på bildebehandlingsapplikasjoner. MONAI følger etablerte fremgangsmåter for programvareutvikling og tilbyr et brukervennlig, godt dokumentert og grundig testet programvarerammeverk som er tilgjengelig for alle interesserte forskere via https://monai.io/.

I denne oppgaven presenterer vi fastMONAI, et lavkode Python-basert dyp læringrammeverk med åpen kildekode bygget på rammeverket fra MONAI. Biblioteket kombinerer flere moderne teknikker ved å integrere funksjoner fra MONAI med to andre kraftige biblioteker: fastai og TorchIO, sammen med skreddersydde moduler.

fastMONAI tilbyr et høynivå-API som forenkler prosessen med datainnlasting, preprosessering, trening og tolkning av resultater, slik at forskere kan bruke mindre tid på koding og fokusere mer på særegne utfordringer i hvert prosjekt. Til tross

for et høynivå grensesnitt, opprettholder fastMONAI muligheter for tilpasning og tilbyr samme fleksibilitet som fastai. Erfarne utviklere kan derfor innlemme tilpassede utvidelser ved behov.

Utviklingen og evalueringen av fastMONAI har blitt utført ved bruk av både offentlig tilgjengelige og kliniske studiedata samlet inn fra flere pasientgrupper, radiologiske domener og organer, inkludert identifikasjon av hjerne og omliggende vev og strukturer (Paper B), lungekreft (Paper C), gynekologisk kreft (Paper D) og ryggsmerter (Paper E). Hver av disse pasientgruppene krever nøyaktig og effektiv medisinsk bildeanalyse for diagnose og behandlingsplanlegging. Resultatene våre i denne oppgaven viser lovende forbedringer i diagnostisk nøyaktighet og mer strømlinjeformet arbeidsflyt.

For å grundig evaluere dyplæringsmodeller, er det imidlertid avgjørende å integrere dem i arbeidsflyt slik de er i den virkelige verden og studere ytelsen deres i realistiske sammenhenger. I denne oppgaven fant vi at fleksibiliteten og det brukervennlige API-et til fastMONAI letter integreringen av trente modeller i klinisk infrastruktur (se figur 4.5). Dette utforskes videre i pågående og fremtidig arbeid som bygger på avhandlingens resultater.

Contents

Part I

OVERVIEW

We are still in the earliest days of deep learning in medicine, facing many challenges and opportunities as the field evolves and matures.

—ChatGPT (Jan 30 version)

CHAPTER 1

DEEP RADIOLOGY: AN INTRODUCTION

1.1 Background

Artificial intelligence (AI) has become a familiar term in today's society, consistently growing in popularity; see figure 1.1. John McCarthy, who coined the term in 1955, described artificial intelligence as *"the science and engineering of making intelligent machines"* [136], and Russell and Norvig as *"the study of agents that receive percepts from the environment and perform actions"* [197]. It is a vast field encompassing many technologies and approaches, from cognitive sciences to search algorithms, knowledge-based agents, and decision theory.

Machine learning (ML) is a subfield of artificial intelligence focusing on mathematical models to recognize patterns in data or, in other words, models that can learn from experience. Deep learning (DL) is a subfield of machine learning that is, in general, based on a specific class of models called artificial neural networks [65]. Deep learning has seen significant growth and advancements over the past decade [256], and it is the main driver for the current interest in artificial intelligence.

One of the key advantages of deep learning is its ability to tackle highly complex tasks such as image recognition and generation, natural language processing, speech recognition and synthesis, and more (see Chapter 2). Traditional machine learning models often rely on manual feature engineering, using domain knowledge to construct suitable transformations of the original features. This can be a time-consuming and challenging task for real-world problems. In contrast, deep learning models are *representation learners*, able to extract useful features directly from raw–or close to raw–input data while learning to solve a given task. They do this by employing models known as artificial neural networks (ANN), consisting of multiple layers of computational units loosely analogous to biological neurons, allowing them to learn patterns in data through a hierarchical learning process. This is explained in detail in Chapter 2.

Different types of ANN models have been successfully applied in various fields. They have recently captured the public's attention through applications in computer vision (e.g., DALL·E 2 [183] and Stable Diffusion [190]), natural language processing (e.g., InstructGPT [164], ChatGPT, GPT-4 [163], and Bard [223]), and speech recognition (e.g., Whisper [180] and AssemblyAI's Conformer-1). Although artificial neural networks have been around for several decades [197], they have gained widespread recognition as one of the best machine learning approaches to a variety of problems in recent years. The breakthrough and continued rapid development of deep learning are mainly due to the availability of open-source projects, new techniques, large annotated public datasets, and increased available computational power.

Deep learning has great potential in medical research and clinical practice and is behind a possible transformation of the healthcare system [182, 225]. This is illustrated by a large number of studies from a wide range of clinical areas, such as stroke diagnostics [58], chest X-ray [31], dermatology [177], and in the analysis of electronic medical records [210]. In particular, deep learning has shown to be well-suited to tasks within image diagnostics and image analysis [129]. The goal that drives deep learning research for medical applications is the development of tools that can enhance the accuracy and efficiency of diagnosis, reduce medical costs, streamline, and improve diagnostic processes through a greater degree of precision medicine, with better prognostics and stratification of therapy [182]. This integration of deep learning in medicine is called "deep medicine" in this thesis, following Eric Topol [224].

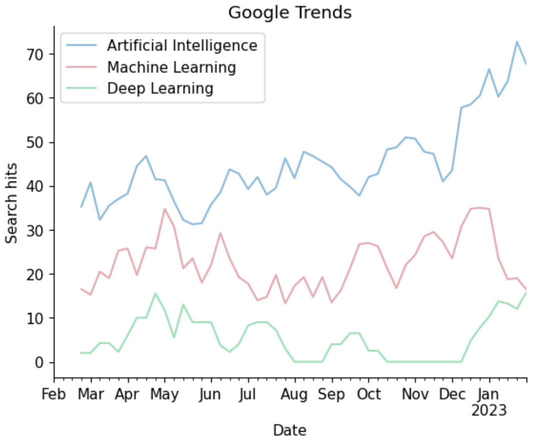


Fig. 1.1: Artificial intelligence is often incorrectly and somewhat misleadingly used as a synonym for machine learning and sometimes even for deep learning. The figure shows the Google Trends chart of searches for the terms "artificial intelligence," "machine learning," and "deep learning" over the past year. Note that the y-axis shows relative popularity with the maximum value scaled to 100. Data provided on GitHub: `https://github.com/skaliy/thesis_supplementary_materials`

Recent developments of large-scale *foundation models* [25] and multi-modal models (handling both image and text inputs) make the field poised to conquer broader swathes of medicine, bringing us closer to the ambition of deep medicine [1, 114, 150][1]. The forthcoming book [115] by Lee et al. will likely provide valuable perspectives on the opportunities and challenges multi-modal models like GPT-4 present for medicine.

Paralleling the developments in deep learning, the past two decades have seen advances in medical imaging technologies that have transformed the field of radiology, leading to increasing use of radiological examinations for the purpose of diagnosis and follow-up of various diseases [15, 209]. This brings with it several challenges. As an example close to home, the Norwegian newspaper Bergens Tidende (BT) reported in 2016 that the Radiology department at Haukeland University Hospital had a backlog of over 7.000 examinations due to a shortage of radiologists [169]. Dagens Medisin reported in 2017 that over a four-year period, the reporting system of the Norwegian Directorate of Health received 203 reports of unwanted events (e.g., deaths, worsening of

[1] One week before the submission of this thesis, Segment Anything Model (SAM) [105], a promptable segmentation model, demonstrated remarkable zero-shot generalization capabilities for 2D images. It will be interesting to explore this model's usefulness for medical imaging and how similar medical imaging-tailored models can be constructed.

disease, over- or under-treatment), many of them directly or indirectly due to a shortage of radiologists [156]. Diagnostic errors were most often caused by misinterpretation, delayed diagnosis, and pathologic findings not noticed by the radiologists. For example, Lauritzen et al. [111] conducted a prospective study involving 1071 CT examinations, where they found that double reading resulted in clinically important changes in 14% of the cases. The shortage of radiologists is a worldwide phenomenon. For instance, at Canberra Hospital in Australia, previous scans reportedly were ignored, and radiologists were working unsupervised due to labor shortages [219]. In 2016 in the UK, most radiology departments paid radiologists to work overtime to cover the backlog [187]. Since that time, the situation has not improved. The latest Royal College of Radiologists radiology census report shows that NHS needs approximately 2000 additional radiologists to reduce the backlog of unread examinations [128]. These global shortfalls are a cause of concern and directly impact patient safety while also causing burnout among radiologists [255].

The limited resources in radiology emphasize the need for new technologies, and there's a hope that deep learning-based applications integrated into clinical workflows can alleviate the issue. This is, together with the hope of improved diagnostics and prognostics, the main reason for the immense interest in artificial intelligence in radiology and other imaging diagnostic sectors and in medicine and healthcare more broadly [149, 182, 225, 230].

It is clear that deep learning applications have the potential to mitigate the burden on radiologists in tasks related to image analysis and provide support to make better decisions. However, there has been a lot of hype surrounding the potential use of deep learning for medical image analysis in recent years, based on results from retrospective single-site studies [182]. Although deep learning models can be accurate during the development phase, a major problem with this approach is that it fails to evaluate how these tools will be used in production, while it is known that there can be a drastic reduction in performance when such tools encounter real-world data [4, 104]. There are few deep learning-based systems implemented in real-world clinical care [240] with demonstrated clinical impact. For instance, in 2021, Leeuwen et al. looked at 100 CE-marked AI products from 54 different vendors and found that only 18 of these products had the potential to produce clinical impact [231].

As Rajpurkar et al. [182] pointed out, standards for transparency in reporting and validation are needed to build trust in deep learning-based systems. Sharing data from clinical studies for research purposes can be difficult due to privacy, ethical, and legal concerns that need to be taken into account [18, 232, 246].

However, sharing code and extensive documentation (both code and data) can greatly contribute to research while protecting patient privacy. Unfortunately, many researchers in the deep learning field focus on publishing their findings and neglect to share the code, data, and documentation necessary for others to reproduce their work [67, 122, 151]. Lack of transparency makes it difficult for others to verify the validity of the results, validate with their own data, and build upon the research to make advances in the field and deliver impact in real-world clinical care [67]. In addition, the field may suffer from the widespread problem of publication bias [207], where positive results are more likely to be favored for publication over negative results [233]. In general, it is a problem if the incentives are towards optimizing for publication (intentionally or unintentionally) rather than the research's real (and advertised) goals.

Most deep learning applications for medical image analysis are developed through supervised learning, models trained based on input data with labels that describe the desired output (e.g., malignant or benign tumor). To develop and evaluate such models, significant amounts of accurate and diverse annotated data are required. Unfortunately, obtaining annotated data in the medical field is often a complex, time- and cost-consuming, and partly an unreliable process (e.g., high inter- and intra-operator variability), particularly for 3D medical images. As a result, a number of important areas within medical imaging and imaging diagnostics have not yet experienced the impact of the latest developments in deep learning. This is because most researchers focus on areas where large amounts of annotated data are already available or where the annotation process is reasonable to carry out, which doesn't necessarily capture the areas of the greatest importance and surely doesn't capture all areas [233]. It is, therefore, important to develop new methods and strategies that enable effective learning from limited amounts of data.

Three main factors impacting the performance of a deep learning system are the network architecture, training methods, and data used for training [244]. Over the past decade, lots of efforts in the field have focused on improving the performance of benchmark datasets with a model-centric approach, looking at developing new model architectures and training strategies. However, benchmark datasets often undergo filtering and cleaning processes, failing to represent the real world [11]. Real-world data are usually more complex, containing a number of data quality issues [199]. In other words, the performance of deep learning systems based on supervised learning is limited by the quality of annotated data, as human decisions serve as the standard of truth. The well-known phrase "garbage in, garbage out," referring to how incorrect input will produce faulty outputs, applies.

A framework championed by, among others, Andrew Ng [153], is to focus on *data-centric development*, a discipline of systematically engineering the data used to train and evaluate the models. A practical demonstration of the value of such an approach in medical imaging is the nnU-Net ("no-new-Net") framework developed by Isensee et al. in 2021 [91], which uses a model architecture introduced way back in 2015 (discussed in Chapter 2) while still achieving top performance in a wide range of medical image segmentation tasks.

To be able to focus on exploratory work and evaluate the quality of the data, you need frameworks for rapid development using existing methods for various tasks rather than "reinventing the wheel" for every project. As the field of deep learning evolves in the medical domain, the community needs practical tools and recipes to help researchers develop effective methods.

Recently, low-code deep learning frameworks, such as nnU-net [91] and Auto3DSeg [33], have been developed to offer practical tools that enable researchers to spend less time on coding and more time on experimentation. By providing a high-level abstraction, these tools may accelerate the development process and make it more accessible to a broader range of users, as developers do not need deep insight into the underlying platform [116].

Moreover, computational notebooks, such as Jupyter [106], have become increasingly popular among machine learning researchers to present research results, as text, illustrations, and executable code can be shared in the same document [172].

While computational notebooks and low-code libraries hold great potential, it is important to mention that they also present challenges that can adversely affect the quality and maintainability of the code and hinder reproducibility. For instance, computational notebooks have been criticized for bad practices in naming files, versioning, testing, and modularizing code [172]. Furthermore, in the case of low-code libraries designed to facilitate rapid software development, complexity can quickly grow, making the resulting software challenging to understand and maintain [116].

To ensure the quality and longevity of research conducted using low-code platforms and computational notebooks, it is essential to support modern software engineering practices (e.g., unit testing, documentation, modular design, etc.). These practices not only improve the reproducibility and reliability of research but also make it easier for other researchers to understand, reuse, and build upon the work.

This thesis aims to contribute in this context by presenting fastMONAI [96], a low-code Python-based open-source deep learning library we have built on top of fastai [79, 80], MONAI [33], and TorchIO [170]. We created fastMONAI to simplify the use of modern deep learning techniques in 3D medical image analysis for solving classification, regression, and segmentation tasks while still giving the users easy access to lower-level functionality. fastMONAI provides users with functionalities to step through data loading, preprocessing, training, and result interpretations.

We have created a set of tutorials showcasing the features and use cases of the library. These tutorials and the fact that the library is low-code help ease the entry into deep learning for medical imaging, making it more accessible to a broader range of researchers, practitioners, and even clinicians.

To approach the ambitions of *deep radiology*–the integration and application of deep learning at the heart of radiology–we need to integrate deep learning-based applications into the clinical infrastructure and examine both the potential challenges and opportunities in real-world settings. By doing so, we can better understand the limitations of current methods, harness their potential, and work to construct the next generation of methods.

This thesis aimed to contribute to the realization of this vision.

1.2 Research and motivation

Throughout the work reported in this thesis, the author has had the opportunity to collaborate closely with radiologists, clinicians, and experts on healthcare IT infrastructure at the Mohn Medical Imaging and Visualization Center (MMIV), which is part of the Department of Radiology at Haukeland University Hospital. The center's vision is to improve decision-making and patient care by developing new quantitive methods for medical imaging in preclinical and clinical settings, which aligns well with the objectives of the thesis work. Our work on the deep learning software framework fastMONAI [96] has served as an organizing principle throughout the thesis work. The aim was to contribute to further developing and evaluating deep learning in medical imaging. The fastMONAI framework was developed through applications based on both public and clinical study data and involving multiple patient groups, including lung cancer [98], gynecological cancer [75], and low back pain [94]. For all of these groups, early detection and providing optimal therapy are crucial for a better quality of life. Although different organs and pathological processes are involved, there are compelling reasons to address these disease groups in the same project:

i) **Quantitivate in vivo imaging** techniques, such as magnetic resonance imaging (MRI) and computerized tomography (CT), are important tools, as they enable clinicians to perform a noninvasive examination of internal body structures. These imaging techniques provide valuable information for diagnosing, planning, and monitoring various medical conditions. For more information, see Chapter 3.

ii) For each organ, the challenge in **data analysis is similar**. For example, the delineation of anatomical regions for analysis and extraction of regions of interest (ROIs) that have diagnostic and predictive value.

iii The components used in deep learning for computer vision projects tend to be quite similar from project to project (see Chapter 2). A **common framework** with reusable program components that can be applied for various organs allows for more efficient development processes, as researchers can focus on the challenges within each project.

iv By facilitating and championing **open science**, we hope to contribute to more widespread sharing of source code in the field of medical AI and in medical imaging more broadly.

1.3 Thesis outline

The thesis is structured into two parts (I and II). Part I consists of four chapters that are organized as follows:

2. Deep learning in computer vision: A picture is worth a thousand layers

This chapter provides a brief overview of computer vision, followed by an introduction to deep learning and its impact on the field over the past decade. This exploration sheds light on the ongoing progress in this rapidly evolving area of research.

3. The domain: radiology and imaging diagnostics

This chapter introduces the fundamentals of diagnostic radiology and imaging methods alongside an exploration of the tools and technologies employed throughout the development of this thesis.

4. From pixel to patient: conclusions, contributions, and continuations

In this chapter, we discuss important challenges and opportunities in the field of diagnostic radiology. Along the way, we look at the contributions of the thesis, and point to possible future work.

5. Summary of papers

This chapter presents a summary of the various papers and research contributions made during the development of this thesis.

Part II consists of one manuscript and four published papers [75, 94–96, 98], and is organized as follows:

A. fastMONAI: a low-code deep learning library for medical image analysis

B. 2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using fastai

C. Pulmonary nodule classification in lung cancer from 3D thoracic CT scans using fastai and MONAI

D. Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer

E. Multi-Center CNN-based spine segmentation from T2w MRI using small amounts of data

The field of deep learning in computer vision offers exciting opportunities for innovation and breakthroughs, but it also requires careful consideration of the challenges it presents in order to fully realize its potential.

—ChatGPT (Jan 30 version)

CHAPTER 2

DEEP LEARNING IN COMPUTER VISION: A PICTURE IS WORTH A THOUSAND LAYERS

In this chapter, we will provide an introduction to the field of computer vision (CV) and various core tasks within this field. Computer vision is a vast field with a long history and consists of a range of approaches and algorithms. However, deep learning has caused a major revolution in computer vision over the past decade, and it is now the go-to approach for many core challenges in computer vision. After a broader look at computer vision, the chapter briefly introduces deep learning before ending with a review of some recent developments in deep learning for computer vision.

2.1 Computer vision

For centuries, researchers have been interested in visual perception, the ability of humans and animals to interpret their surrounding environments from the information received by the retina [133]. Biological visual perception systems are full of interesting, evolved features and characteristics. For instance, frogs have a "bug perceiver" cell in their retina known as *net convexity detection* cells that are capable of detecting small dark objects [117]. Wiesel and Hubel had a number of groundbreaking discoveries on how the visual systems work in cats and monkeys [85–87].

Computer vision (CV) is a subfield of artificial intelligence (AI) that seeks to replicate biological visual perception in machines. It involves the use of mathematical techniques to perceive and understand objects in images and videos [217]. Early attempts in image understanding involved, among other techniques, edge detection [46], shape from shading [76], stereo matching [134], image segmentation by tree traversal [78], and optical flow [77].

Computer vision has undergone explosive growth over the recent decade (see figure 2.1), caused by breakthroughs in deep learning that were achieved mainly due to advancements in parallel processing on graphical processing units (GPUs) [166], GPU-enabled deep learning models [108], and access to large labeled datasets (e.g., ImageNet [48], Microsoft COCO [121], etc.). These developments have resulted in significant advancements in terms of performance and reliability in various domains, including robotics [101], remote sensing [130], medical imaging [38, 129], and many more [256].

Despite these developments, models based on deep learning are still error-prone and can't compare to human visual perception and our ability to generalize. For instance, such models can be biased in subtle ways: De Vries et al. [47] looked at six publicly available object-recognition systems for household items and discovered a decline in their performance when presented with data from low-income countries. This was partly due to the appearance of objects in a different setting from the data it was trained and evaluated on (e.g., toothbrushes outside bathrooms). As we can see from this example, deep learning models lack the ability to learn abstractions as humans do.

Computer vision is, in nature, an inverse problem, i.e., recovering information about the three-dimensional structure of the world from images, including shape, color, and illumination [217]. As discussed in chapter 1, most deep learning systems are developed using supervised learning (see section 2.3). These models' performance is highly dependent on the data used to train and evaluate them in terms of both quality and quantity (see Chapter 4). As noted by Marcus [132], deep learning models are, at the core, just statistical methods, and like any statistical method, they suffer from deviation from their underlying assumptions.

Deep learning models have great potential for solving complex tasks in a wide range of domains. However, to ensure accurate and reliable tools, clear reporting standards are needed to understand the limitations and capabilities, as highlighted by Riley [186].

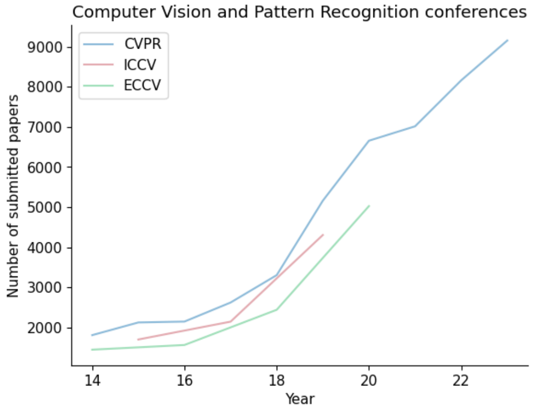


Fig. 2.1: Number of submitted papers for computer vision and pattern recognition conferences in the last decade. Source: https://github.com/lixin4ever/Conference-Acceptance-Rate

2.2 Computer vision tasks

Humans can naturally perform various visual tasks without explicit instruction, which stands in contrast to computer vision systems. In computer vision, we need to precisely define the task to solve, and often we need to split a task into sub-tasks using individual modules, each with explicit instructions. For this reason, there is a wide range of computer vision tasks, and in this section, we explain three commonly used ones: image classification, object detection, and semantic segmentation.

2.2.1 *Image classification*

Image classification refers to the ability of machines to categorize what they see in images.

This task can be divided into two main cases [197]. In the first case, identify a single object in an image, as illustrated in figure 2.2 **a**, where the image has been labeled as "defected." Or figure 2.5, where we want to distinguish between blue and green bottles.

In the second case of classification, the task is to identify multiple objects in an image (known as multi-label classification). For instance, if an image contains both blue and green bottles, the task is to label both of them correctly.

2.2.2 Object detection

Object detection, on the other hand, involves both the classification and localization of objects. As shown in figure 2.2 **b**, this is usually done by using bounding boxes that define the object(s) coordinates within an image. For an up-to-date survey of object detection, see [263].

2.2.3 Semantic segmentation

Semantic segmentation is a challenging task that involves assigning a label to each pixel (or voxel in 3D) within an image based on a pre-defined set of classes. The goal is to accurately outline the shape of the object(s) of interest. Figure 2.2 **c** shows an example of semantic segmentation. The boundary of the bottle was manually outlined by the author.

Medical image segmentation involves precise delineation of regions of interest (e.g., organs and tumors) from their surroundings in image modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), see Chapter 3. Segmented objects provide valuable insight into the characterization of the region of interest (e.g., tumor volume, shape, location, etc.) [64], and can be described as the "holy grail" of medical image analysis [129].

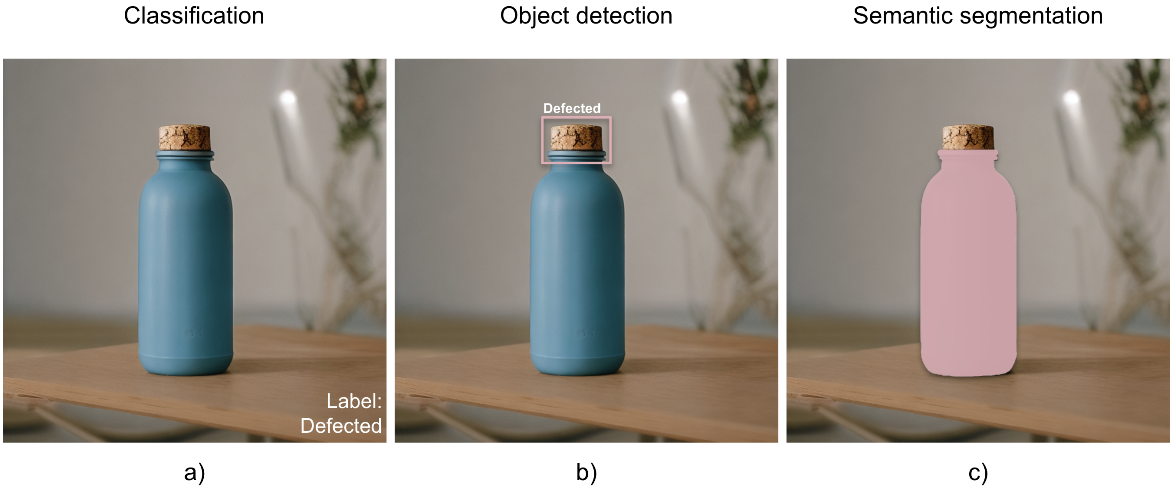


Fig. 2.2: Illustration of various tasks in computer vision: **(a)** classification, **(b)** object detection, and **(c)** semantic segmentation. The image is generated by the author using Stable Diffusion 2.1. The source code to reproduce the original image is available on GitHub: `https://github.com/skaliy/thesis_supplementary_materials`.

2.3 Solving vision tasks by learning

Machine learning models learn from data and make decisions based on the acquired knowledge without being explicitly programmed [66]. There are several ways these models can learn. This section will focus on **supervised learning**, the learning approach used in all the projects presented in this thesis.

As the name implies, supervised learning involves machine learning models that learn under the guidance of feedback. A common supervised learning task is classification, a topic already covered in section 2.2. Another typical task is solving regression problems, the task of predicting continuous values, such as the bounding box coordinates in figure 2.2 **b**.

The general workflow of supervised learning is shown in figure 2.3. For a given task, a machine learning model is presented with a number of labeled samples in the form of input data $\mathbf{X}$ with corresponding output labels $\mathbf{Y}$, referred to as training data. During the learning process, the machine learning model learns rules for the mapping function[1] $\mathbf{X} \rightarrow \mathbf{Y}$ by iteratively minimizing the difference between the predicted output $\hat{\mathbf{Y}}$ and the true label $\mathbf{Y}$ using loss function (also known as cost function, see table 2.1). These rules–for a given machine learning model–are typically encoded in the *parameters* θ of the model[2]. Mathematically, we can formulate it as follows:

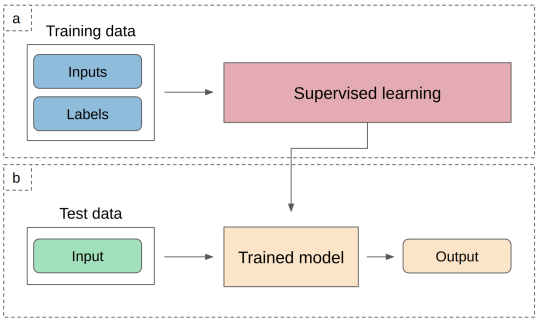


Fig. 2.3: A simple overview of supervised learning workflow. **(a)** An appropriate model is trained to map inputs to outputs by minimizing the loss function. **(b)** The trained model is then applied to new, unseen data.

$$\hat{\mathbf{Y}} = f(\mathbf{X}, \theta)$$

When the machine learning model has been trained, the goal is to accurately predict the output $\hat{\mathbf{Y}}$ for new, unseen inputs $\mathbf{X}$, referred to as test data. In practice, to get an unbiased evaluation of how well the model generalizes, practitioners usually set aside a portion of the labeled data, for example, 20% of the data, for testing. However, the specific percentage may vary based on the amount of data you have available and the complexity of the task. The work needed to obtain labels also depends on the task. For instance, if you think about the task of labeling the bottle in figure 2.2, the labeling process for image classification is quicker than for object detection, which in turn is faster than the manual annotating process for semantic segmentation.

In addition to testing, people usually set aside a portion of the training data, known as validation data, used to select the appropriate model for the given task. An alternative approach is to partition the training data into k subsets, known as k-fold cross-validation. This process is more computationally expensive and not

[1] Note that, for simplicity, we assume the existence of a functional relationship between the inputs and outputs.

[2] Here, we exclusively discuss parametric models.

always possible. A limiting factor with supervised learning is that the accuracy of the predictions is highly dependent on the quality and representativeness of the labeled data, a topic discussed in Chapter 4.

Some common machine learning models for supervised learning include linear regression, support vector machines, decision trees, random forests, gradient boosting, and artificial neural networks. Some of these models are restricted to either classification or regression, like linear regression. Others, like neural networks that we will discuss shortly, can be used for both tasks.

2.4 Deep supervised learning in computer vision

Artificial neural networks (ANNs) are a specific family of machine learning models that form the core of deep learning, and they are typically trained using supervised learning [3]. Neural networks are not a recent invention; the concept has been around for several decades. They were first introduced by McCulloch and Pitts [138] back in 1943, where they presented a computational unit inspired by biological neurons, which later became known as an artificial neuron. Since then, many scientists have been contributing to this field: Minsky and Edmond's SNARC from 1951 [197], Rosenblatt and his perceptron [192], Minsky and Papert's influential book on the limitations of perceptrons [145], Fukushima and the neocognitron [61], Rumelhart et al. work on learning representations by backpropagation algorithm [195], LeCun et al.'s CNN trained by backpropagation (later known as LeNet) [113], Deep Belief Network proposed by Hinton et al. [74], and many more. See [202] for a historical overview.

The basic concept of an artificial neuron, which is the computational unit in artificial neural networks, is depicted in figure 2.4 **a**.

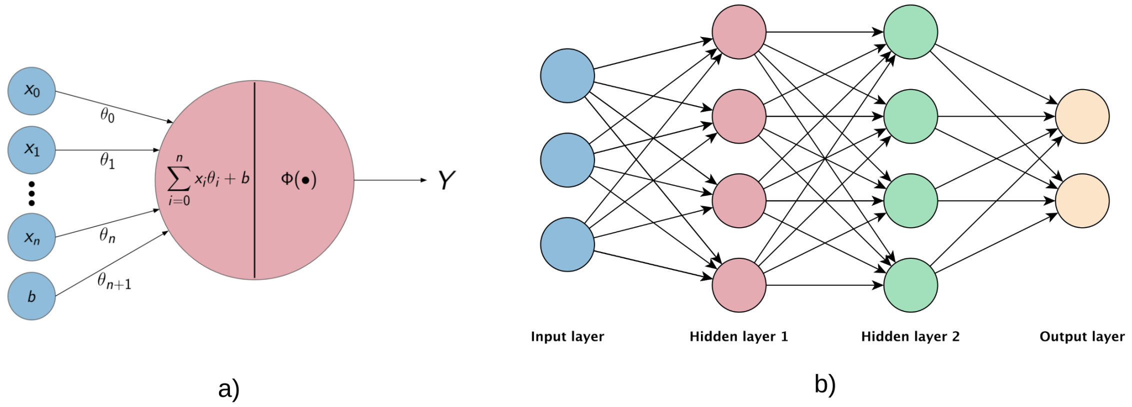


Fig. 2.4: Figure **(a)** shows the basic concept of an artificial neuron. First, the dot product between the input data **X** and its corresponding parameters θ, called weights, is calculated. The value is then passed through an activation function Φ to decide its output (see table 2.1). Figure **(b)** illustrates a deep neural network with two hidden layers. In this setup, all the neurons within a layer receive input from all units from the previous layer.

Modern neural networks are composed of multiple artificial neurons stacked together in layers. They typically consist of an input layer, one or more hidden layers, and an output layer, as shown in figure 2.4 **b**. Neural networks with a single hidden layer are called shallow networks, while networks with multiple hidden layers are called deep neural networks (hence the term "deep" in deep learning). Choosing the optimal number of layers and neurons is a difficult task that highly depends on the complexity of a given task and the number of training samples available for the learning process [102]. A deep neural network can represent a higher level of complexity by increasing the number of layers and units within each layer [65]. However, having an excessive number of layers may result in *overfitting*-a phenomenon where models become overly adapted to the training data and fail to generalize to new, unseen data.

To avoid overfitting, many methods can be applied. Collecting more data would often help, but this is often practically impossible. Other options are to decrease the complexity by reducing the number of hidden layers or/and applying various regularization techniques (see table 2.1). However, before employing these methods, one should confirm overfitting during the training process [79]. A general practice to detect overfitting is to set aside validation data (see section 2.3) and monitor if, at any point during the training process, the validation performance gets worse while the training performance continues to improve. If so, it is an indication that the model is overfitting.

The basic steps to train any deep learning models in a supervised fashion are as follows:

```
1. Initialize the weights, either randomly or using weights from
an already trained model (pre-training)
2. Get predictions from a batch of data by feeding it to the model
3. Calculate the performance using a loss function
4. Calculate the gradient of the loss with respect to each parameter
using backpropagation. These gradients measure how changing each weight
would change the loss.
5. Move each weight a little bit in the opposite direction of
its gradient using gradient descent with a specified step size (called
the learning rate)
6. Go back to step 2 and repeat the process
7. Do this over and over until you meet some chosen stopping criteria
```

These seven steps of training a neural network are illustrated in Fig. 2.5**a**.

One of the benefits of neural networks is their capacity to automatically learn useful representations from raw data without human interaction [112]. In the past decade, different variants of deep neural networks have become the methods of choice for a wide range of computer vision tasks. Image classification is a prime example. In 2012, Krizhevsky et al. [108] presented "AlexNet", a convolutional neural network (CNN) [61, 113] that surpassed previous methods on the ImageNet dataset [48]. Since then, various CNN architectures have attracted widespread attention and are currently the core elements in many computer vision tasks (see section 2.5). CNNs play a crucial role in all the work reported in this thesis, and therefore, we provide a short summary of the main ideas behind CNNs as illustrated in figure 2.6.

Convolutional neural networks often consist of convolutional, pooling, and fully connected layers. As the name implies, convolutional layers are the fundamental building blocks of convolutional neural networks. These consist of a number of

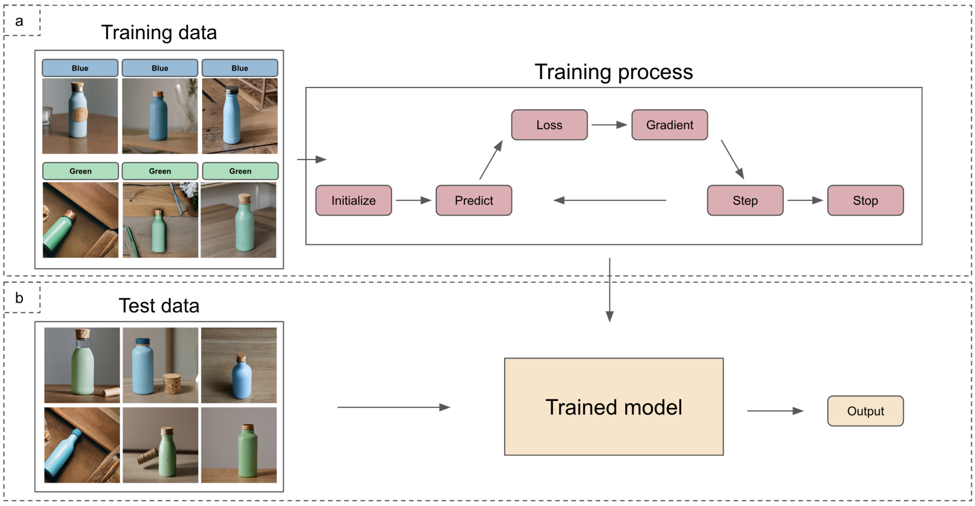


Fig. 2.5: (a) The training process for any deep learning models (adapted from Howard and Gugger [79]), illustrated using a classification task of identifying green and blue bottles. **(b)** The trained model is used to run inference on new, unseen images.

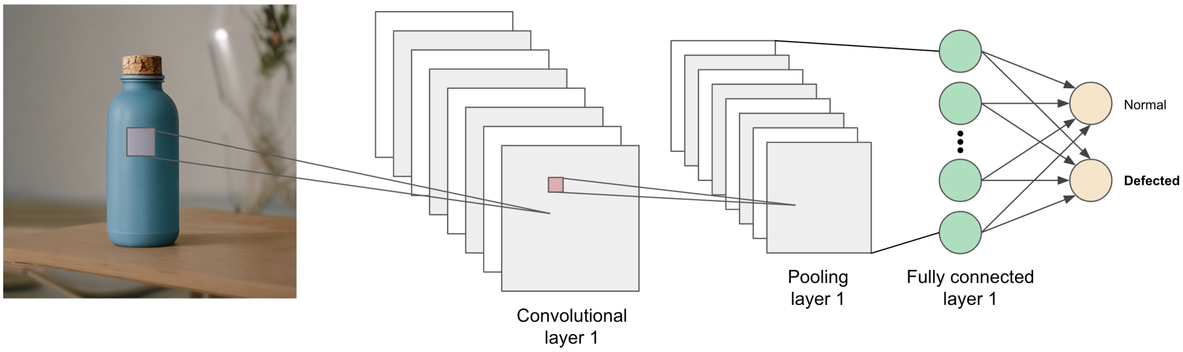


Fig. 2.6: An illustration demonstrating the information flow in a basic CNN architecture, comprised of an input layer, a convolutional layer, a pooling layer, and a fully-connected layer, designed to detect defective bottles in images.

learnable kernels (also called filters) that slide across an input tensor (e.g., 1D, 2D, 3D) to calculate the dot product between the kernels and the inputs. This results in outputs that form what are called activation maps (also known as feature maps). The sliding size is determined by a hyperparameter called *stride*. Note that each neuron within a convolutional layer is only connected to a specific region in the previous layer (the local receptive field), enabling them to capture spatial information within an image and automatically extract relevant features. As shown by Zeiler and Fergus [254], the filters in the first layer of simple CNNs typically learn to detect simple features such as edges, color, etc. As we move to deeper layers, they will learn to extract complex task-specific features. The pooling layers are used to downsample activation maps in order to reduce the number of parameters in the model, lowering computational requirements and the risk of overfitting. Popular pooling methods include average pooling [113] and max pooling [184], where the average and maximum value from a pooling window is calculated, respectively. Note that these layers do not have any learnable parameters, which can lead to the loss of important information. For this reason, many architectures include strided convolutions (i.e., increase the sliding size to reduce the activation map), such as the ResNet model [72].

Fully connected layers can be added at the end of CNN architectures for classification and regression tasks. These layers map the extracted high-level features from the final convolutional or pooling layer into output for a given task [3]. As we can see in figure 2.6, these layers work the same way as the feed-forward network shown in figure 2.4. If fully connected layers are included in a CNN, it should be noted that they can easily end up containing most of the parameters in the network [14].

Table 2.1 provides a glossary of some of the main concepts. Comprehensive textbooks on deep learning include [3, 65, 197].

Table 2.1: Glossary of core terms, techniques, and concepts in deep learning.

Activation function	A non-linear function applied after calculating the weighted sum of inputs in a neuron to decide its output. Some commonly used activation functions include Sigmoid, Tanh, ReLU, and Leaky ReLU. [159].
Architecture	Feedforward neural networks, convolutional neural networks (CNNs) [61, 113], generative adversarial networks (GANs) [66], transformers [234], etc. See the review paper provided by Chai [35] et al. for more architectures.
Backpropagation	Calculate the gradients of the loss function with respect to all the weights in the model.
Fully convolutional network	A model composed of only convolutions and pooling layers, without any fully-connected layers, thus reducing the number of parameters. Typically used for semantic-segmentation [127] and object detection [45] tasks.
Loss function (cost function)	Evaluates the difference between predicted output and the ground truth, providing a measure of how well the model is performing on a given task. The primary goal during model training is to minimize the loss function.
Model	Consists of a design architecture and a specific set of parameters to map input data to output labels.
Optimizer	Used to change the parameters during training to minimize the loss. Gradient descent and variants of this algorithm are the most common way to optimize neural networks [193].

[3] Traditional machine learning models can also be applied directly to the output features from the final convolutional or pooling layer.

Parameters	Weights and biases in neural network models that define the connection between neurons. These parameters are adjusted during the training process to optimize the performance of the model on a given task. In the case of convolutions, the weights are represented as kernels.
Pre-trained model	A model with trained parameters, usually using large datasets (e.g., ImageNet [48], Microsoft COCO [121], etc.).
Regularization	Set of techniques applied during the training process to reduce the risk of overfitting (e.g., data augmentation [205], dropout [211], early stopping [175], etc.).
Representation learning	The ability to automatically learn to extract useful features from raw input data needed for a given task (classification, detection, etc.).
Self-supervised learning	Self-supervised learning is a training technique where we train a model without the need for human-annotated labels by using labels that are part of the input data. For example, feature learning by removing pixels from an image and reconstructing them [168]. In recent years, various self-supervised learning methods have been successfully used for pre-training models in multiple fields, including natural language processing [49, 82] and computer vision [36, 37, 71].
Transfer learning	The ability of a model to reuse the knowledge learned from tasks in a source domain and apply it to a different task in a target domain during the training process.

2.5 Some recent trends in deep learning architectures for computer vision

As we have seen several examples up to now, deep learning has experienced explosive growth in recent years. Numerous CNN architectures have been proposed following the success of "AlexNet" [108]. The generic nature of deep learning methods makes applying them to various domains feasible, and the knowledge and insights acquired from one domain can be transferred to another domain [36] [4]. In table 2.2, we highlight some important and recent developments of deep learning architectures in computer vision for classification and segmentation tasks.

[4] Also illustrated by our work on a remote sensing task [97] in which we used our experience from medical imaging.

Table 2.2: List of recent deep learning architectures and some high-level descriptions. As is clear from their descriptions, the ideas behind the different architectures build upon each other.

ResNet [72] After the success of "AlexNet," deeper models demonstrated higher performance on the ImageNet classification challenge [216]. However, training deep networks is challenging, partly due to the vanishing gradient problem [17, 63]. In deep models, the gradient may become smaller and smaller (close to zero) as it backpropagates through the layers, resulting in poor convergence during training. ResNet was introduced by He et al. to tackle this exact problem using skip connections. These connections create an additional pathway between different weight layers of the network, skipping one or more nonlinear activation functions. The authors demonstrated the performance of ResNet by winning ILSVRC [196] using models with up to 152-layers, as well as in the Microsoft COCO [121] detection and segmentation challenge in 2015. In their work, they also presented an analysis of using models with more than 1000 layers on the CIFAR-10 dataset [107]. This was a major breakthrough. Before this point, the deepest model was only 22 layers deep [216].
ResNeXt [249] An extension of the ResNet architecture with an additional dimension called *cardinality*, a *split-transform-merge* strategy that was introduced in the Inception architecture used by GoogleLeNet [216] (the ILSVRC 2014 winning architecture). This strategy splits the input data into multiple paths, each undergoing a different transformation before being fused into a single output.
U-Net [191] U-Net was developed by Ronneberger et al. in 2015 [191] for biomedical image segmentation. The architecture is a U-shaped, fully convolutional network consisting of two parts: an encoder (the contracting path) and a decoder (the expansive path), with skip connections between them to preserve useful feature information. The encoder part follows the typical CNN architecture shown in figure 2.6 (excluding the fully connected layers), which is used to extract relevant features from images. Nowadays, the encoder part is commonly an ImageNet pre-trained classification model [89]. The decoder part upsamples the extracted features using transposed convolutions (i.e., inverse convolution operation) to reconstruct a segmentation mask with the same dimensions as the input image. To improve the performance of the original U-Net architecture, several adaptations have been proposed, including 3D U-Net [40], attention U-Net [162], residual U-Net [258], and many others [32, 70, 260]. Moreover, in recent years, U-Net has been successfully applied in other domains as well, e.g. [97, 103, 120]. As demonstrated by Kugelman et al. [109], the original architecture is still a strong baseline model.
Vision transformers [54]

The transformer architecture was introduced in 2017 to solve problems in the field of natural language processing (NLP) [234]. The primary function of the proposed encoder-decoder architecture was the self-attention mechanism used to dynamically learn the relationships between different words in a sequence of words, which allows the model to understand long-range dependencies. Well known language models such as BERT [49], LaMDA [223], and various GPT models [27, 181] are built based on this architecture. However, the performance of these models highly depends on pre-training on large-scale datasets using a self-supervised learning approach [49, 181].

The great success of employing transformers in NLP has sparked a growing interest in applying the architecture for computer vision tasks. In 2020, Dosovitskiy et al. introduced Vision Transform (ViT) [54], a convolutional-free model applying the encoder proposed in the original transformer paper [215] on images represented as a sequence of non-overlapping patches (like words) for image classification tasks. Their model pre-trained on JFT-300M dataset [215] (containing 300 million images) demonstrated better performance and efficiency than ResNet-based models. However, the authors reported that ViTs pre-trained on ImageNet (consisting of 1.2 million images) perform worse than CNNs. In addition, the architecture was unsuitable for vision tasks with high-resolution images due to its quadratic computational complexity with respect to the input size. These challenges led to the development of several architectures built upon the work of ViT. This includes DeiT [226] using training strategies for ImageNet pre-training, and Swin Transformer [124, 125], applying hierarchical architecture with a local attention structure, which is reminiscent of convolutions, to serve as a vision backbone for existing models, such as U-Net [191]. An overview of recent developments in vision transformers is provided in [69].

ConvNext [126]

In 2022, Liu et al. [126] set out to explore whether models based entirely on standard convolutional neural network components could compete with Transformers by incorporating modern design decisions from recent work on Transformers. Starting from a basic ResNet-50 model, they progressively added modern model training techniques and network designs, partly based on the ideas of the Swin Transformer [125] and DeiT [226]. The performance of the initial model was greatly improved by just applying the training techniques utilized in Swin Transformer. The performance was further improved by implementing various design methods from Swin Transformer, including replacing ResNet with ResNeXt. Their final model outperformed the Swin Transform on ImageNet [48] classification, Microsoft COCO [121] object detection, and ADEK20K [259] segmentation. In addition, the authors demonstrated that their architectures, dubbed ConvNext, could be as scalable as vision transformers. Recently, Woo et al. [244] introduced ConvNext V2, an architecture optimized for self-supervised learning, resulting in improved results on the datasets reported in the original ConvNext paper[126]. Their work has shown that the jury is still out regarding Transformers versus ConvNets, that ideas from one approach can be incorporated in others, and the importance of fair comparisons when evaluating design choices.

There is a constant influx of new models that outperform–or claim to outperform[5]– the previous state-of-the-art in various benchmarks. After the tremendous progress in the application of deep learning to computer vision starting around a decade ago,

[5] see Chapter 4 for a critical discussion about many such claims

the field seemed for a few years to have settled on a relatively small set of models for most computer vision tasks–for example, variants of ResNets for image classification and variants of U-Nets for image segmentation. However, the recent trend with vision transformers, a completely different network architecture design, combined with self-supervised learning, hints at a possible new acceleration in the field, offering a new set of tools in the toolkit (and new additional challenges to be overcome).

As we come to the end of this chapter, it is worth reflecting on Maslow's famous adage from 1966 - *I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.* [135]. While various deep learning approaches have made significant advances in computer vision tasks (e.g., image classification, detection, and semantic segmentation), traditional computer vision methods still have their place in the field. Some computer vision problems do not require the complexity of deep learning approaches, and more straightforward traditional methods can solve them efficiently, robustly, and directly. As we saw at the beginning of this chapter, techniques such as edge detection, thresholding, and many others have been around for several decades. A key advantage of many traditional methods is that they do not require large amounts of training data and can be more computationally efficient than deep learning models. Therefore, when deciding between traditional methods and deep learning models for solving a computer vision task, it is vital to consider the problem's complexity. When in doubt, always start with traditional methods (especially if you have a limited amount of labeled data)[6].

It is important to note that a hybrid approach that combines traditional computer vision methods with deep learning techniques may be the best solution in some cases. See [165] for a recent exploration of traditional versus deep learning-based methods in computer vision. An area we would like to flag, albeit not explore, is the intriguing ongoing work on combining artificial neural networks with more principled physical and mathematical models, often dubbed *Scientific ML* [178, 179], and work done on forcing artificial neural networks to produce outputs satisfying useful constraints, as exemplified by [123].

[6] The first rule of machine learning is *"Don't be afraid to launch a product without machine learning"* [262].

The application of deep learning in radiology is only as good as the understanding of the field it is based upon. Without sufficient domain knowledge, the results will be superficial at best.

—ChatGPT (Jan 30 version)

CHAPTER 3

THE DOMAIN: RADIOLOGY AND IMAGING DIAGNOSTICS

Diagnostic imaging is a crucial part of modern medicine [53]. A radiologist, a physician specializing in diagnostic imaging, uses various medical imaging techniques to get insight into the structure and function of organs and tissues in the body that can be used to diagnose, treat, and monitor a wide range of medical conditions [19]. Since Roentgen introduced X-rays in 1895, there have been remarkable advancements in medical imaging [214], and many new imaging modalities have been developed, including Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) (see Section 3.1 below). In parallel, a transforming healthcare technology called Picture Archiving and Communication System (PACS) was introduced [39]. PACS systems enabled clinicians to shift from analog films to digital images, resulting in better storage and access to patients' medical images and records [213]. Today, radiologists can import patient studies, interpret them, and make decisions based on the findings in the images using the PACS display workstation component, regardless of their locations [83].

As discussed in Chapter 1, these technological advancements have increased the use of radiological examinations [15, 209]. Unfortunately, the number of practicing radiologists is not growing as fast as the demand for their expertise [128]. This development is a concern as it directly impacts patient safety and may lead to burnout among radiologists [255]. The limited resources of radiologists emphasize the need for new technologies, such as computer-aided diagnosis (CADx). In recent years, tools based on deep learning have shown great promise in radiology and imaging diagnostics [129, 149], but, as described in the previous chapter and will discuss further in Chapter 4, they also come with a significant set of challenges and drawbacks [8, 60, 151, 188, 194, 220, 222, 248].

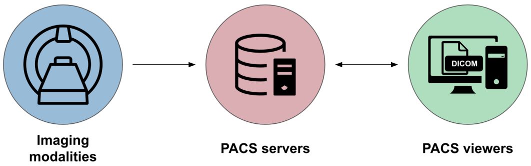


Fig. 3.1: A simplified overview of the diagnostic radiology workflow. Images are recorded by imaging modalities, sent to PACS servers, and inspected by medical specialists using dedicated PACS viewers.

To develop tools that will benefit radiologists, it is essential to understand their existing clinical workflow and the tools they rely on. Figure 3.1 shows a simplified overview of the major components needed in diagnostic radiology workflow, which consists of diagnostic imaging devices (imaging modalities), image archive servers, and display stations. In this chapter, we will have a closer look at these components and examine the tools and technologies used in this thesis.

3.1 Imaging modalities

Imaging modalities refer to various methods used to acquire images of the internal body. Some well-known modalities developed in the last fifty years include CT, MRI, ultrasound, positron emission tomography (PET), and single-photon emission computed tomography (SPECT) [19]. The choice of modality depends on the clinical question and the suspected medical conditions [68, 241], and sometimes multiple imaging modalities may be combined to achieve a more comprehensive assessment [44].

In my thesis, I have exclusively used imaging data from CT and MRI examinations. Therefore, I will limit the discussion to brief explanations of these two imaging modalities.

3.1.1 *Computed tomography (CT)*

A CT scan uses a beam of X-rays around the body and computer processing to create detailed cross-sectional image slices of various internal structures (e.g., bones, blood vessels, and organs) [214]. As a non-invasive method, CT is well-suited for diagnosing many medical conditions. For example, it is the primary imaging modality used for diagnosing patients with lung cancer [176] and acute stroke [174]. Figure 3.2 depicts sections of a lung CT image from the Lung Image Database Consortium image collection (LIDC-IDRI) dataset [10] in three different anatomical planes: axial, sagittal, and coronal.

While CT is considered safe, it is important to note that patients are exposed to a small dose of ionizing radiation during scans. Technological advancements in recent years have led to the development of CT machines that can produce high-quality images while reducing radiation doses to patients. However, as emphasized by McCollough [137], CT scans should only be requested when needed, as with any other medical procedure.

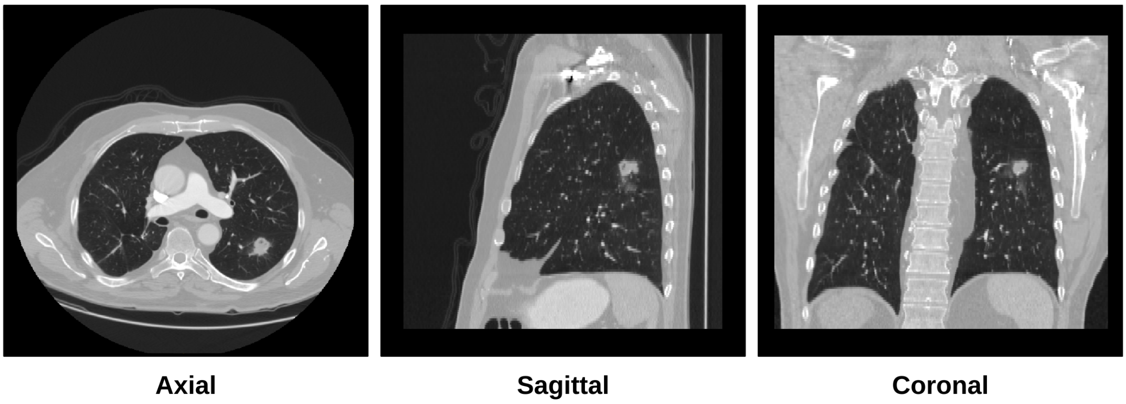


Fig. 3.2: Image slices showing axial, sagittal, and coronal planes of a chest CT from LIDC-IDRI lung nodule dataset [10].

3.1.2 Magnetic resonance imaging (MRI)

MRI technology and scanner availability have also progressed remarkably over the last 20 years [229]. Today, MRI is an essential tool for diagnosing various diseases such as brain tumors [118] and multiple sclerosis (MS) [59]. In contrast to CT, MRI does not expose patients to ionizing radiation. Instead, MRI uses a high-strength magnetic field, radio-frequency pulses, and gradient waveforms to construct images of the internal body. A pulse sequence is a set of instructions determining how radio-frequency pulses and gradient waveforms are used during an MRI scan [20]. In clinical practice, multiple pulse sequences are needed to evaluate different tissues and organs, each designed to highlight various properties (e.g., fat, fluid, etc.). For example, the T1-weighted sequence is a fundamental pulse sequence in MRI, where fat tissue appears bright, while regions with fluid appear dark. Other widely used sequences include T2-weighted, Fluid-Attenuated Inversion Recovery (FLAIR), and contrast-enhanced T1-weighted (T1ce). See Figure 3.3 for an illustration of the different properties highlighted by these sequences.

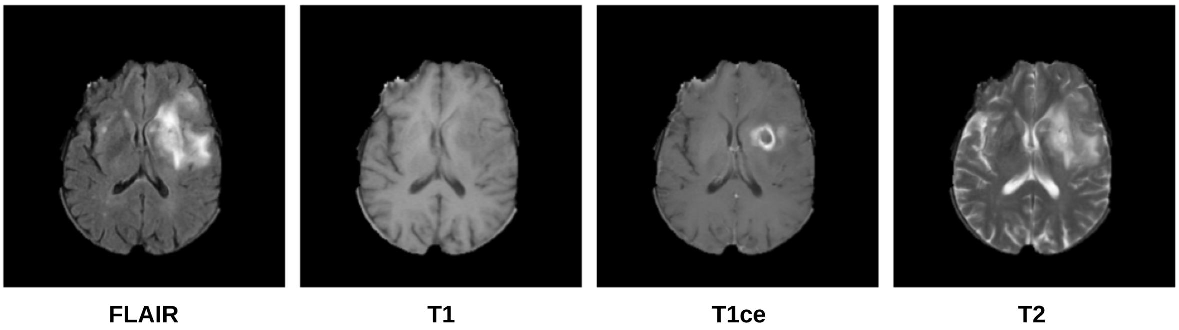


Fig. 3.3: Axial slice samples of FLAIR, T1, T1ce and T2 sequences from the Multimodal Brain Tumor Segmentation Challenge (BraTS) [12, 13, 143] provided in the Medical Segmentation Decathlon challenge [7]. The different sequences highlight various characteristics of the brain tumors located in the left frontoparietal region.

The image resolution and scanning time for a given task depend on the magnetic field strength measured in Tesla (T). With higher magnetic field strengths, the MRI images' signal-to-noise ratio (SNR) increases, resulting in higher-resolution images or faster scanning with the same resolution [158]. As MRI technology has evolved, there has been a trend towards higher magnetic field strengths [218].

Most clinical scanners today have magnetic field strengths of either 1.5T or 3T [201]. In some research and clinical settings, ultra-high-field MRI with magnetic field strengths of 7T or higher is used to achieve even higher resolution [158, 229]. This increase in magnetic field strength allows more detailed visualization of anatomical structures. Figure 3.4 shows an ex-vivo human brain acquired over 100 hours of scan time [55]. Despite the potential advantages, ultra-high-field MRI poses several challenges, including non-uniform radiofrequency fields (e.g., anatomically irrelevant intensity variation) and tissue heating [110].

It is important to note that although MRI is a powerful non-invasive diagnostic tool, it is more time-consuming, expensive, and complex compared to other imaging modalities such as CT [93].

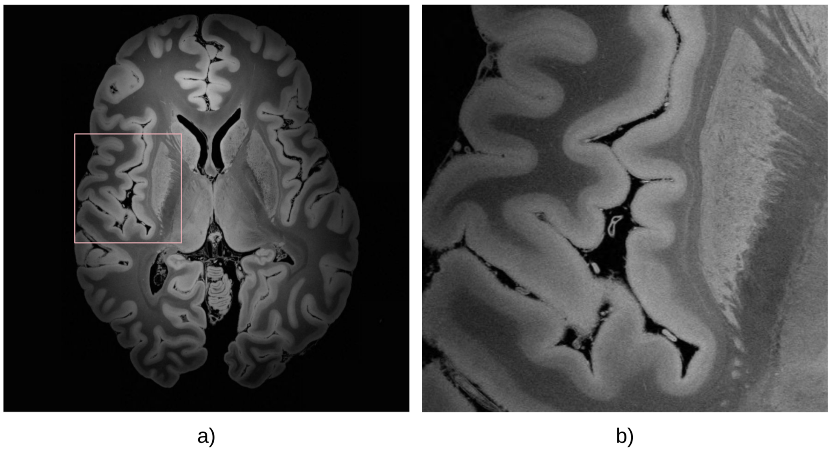


Fig. 3.4: (a) An ex vivo brain MRI of a 58-year-old woman with no history of neurological disease acquired over 100 hours of scan time [55]. **(b)** A magnified view of the highlighted region within the brain showcases the high-resolution details captured by the scanner.

3.2 Digital Imaging and Communications in Medicine (DICOM)

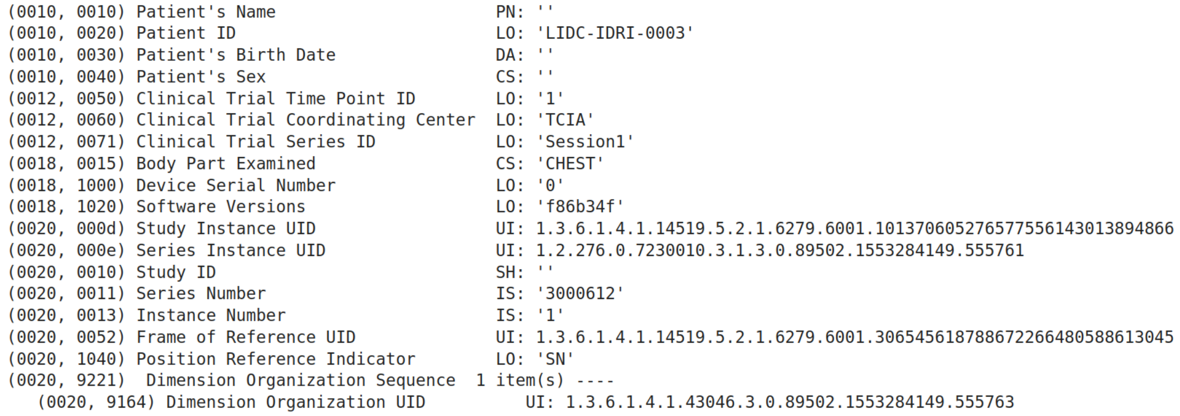
In clinical settings, DICOM is the standard image format and protocol to store and transfer image data and health-related information [22]. Today's digital image acquisition devices, such as CT and MRI scanners, produce DICOM images [171]. DICOM does not only store the pixel data; it also maps related metadata information such as patient information(e.g., name, sex, age, etc.) and image acquisition parameters (slice thickness, pixel spacing, orientation, etc.). The DICOM protocol has entries for more than 2000 elements[1]. Each element is numbered with a unique group and element (tag) and organized in a four-layer hierarchical structure in the following order: patient (Patient ID), study (Study Instance UID), series (Series Instance UID), and image (Image SOP Instance UID). Figure 3.5 shows selected tags, elements, and values from a DICOM file in the LIDC-IDRI dataset [10]. As we can see, some attribute values (e.g., patient name, birth date, sex, etc.) do not have any data, as they have been removed in an anonymization process to protect the sensitive health information about the participants. For research purposes, the DICOM model is extended to include project and scanning event information (DICOM group 0012).

Although DICOM is a powerful and reliable tool, its complexity makes it challenging. In addition, it is not designed for efficient data manipulation and image processing [239]. For example, volumetric images are divided into slices. Each slice is stored as a separate file with its own DICOM tags that provide information about the image and its position and orientation in space. Attempts to store multiple slice data as frames or enhanced DICOM exist but are not widely used and supported across vendors [6]. Consequently, using DICOM for large-scale studies can be cumbersome, as a single visit can contain a large number of individual files.

Due to the complexity of DICOMs, in medical image research, they are often converted to simpler image formats such as Neuroimaging Informatics Technology Initiative (NIfTI) that can store multidimensional data [119]. However, note that these

[1] See https://dicom.nema.org/medical/dicom/current/output/html/part06.html

simpler formats retain only a small set of image metadata from DICOM files, making it challenging to correctly convert them back to DICOMs.



```
(0010, 0010) Patient's Name                      PN: ''
(0010, 0020) Patient ID                          LO: 'LIDC-IDRI-0003'
(0010, 0030) Patient's Birth Date                DA: ''
(0010, 0040) Patient's Sex                       CS: ''
(0012, 0050) Clinical Trial Time Point ID        LO: '1'
(0012, 0060) Clinical Trial Coordinating Center  LO: 'TCIA'
(0012, 0071) Clinical Trial Series ID            LO: 'Session1'
(0018, 0015) Body Part Examined                  CS: 'CHEST'
(0018, 1000) Device Serial Number                LO: '0'
(0018, 1020) Software Versions                   LO: 'f86b34f'
(0020, 000d) Study Instance UID                  UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.101370605276577556143013894866
(0020, 000e) Series Instance UID                 UI: 1.2.276.0.7230010.3.1.3.0.89502.1553284149.555761
(0020, 0010) Study ID                            SH: ''
(0020, 0011) Series Number                       IS: '3000612'
(0020, 0013) Instance Number                     IS: '1'
(0020, 0052) Frame of Reference UID              UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.306545618788672266480588613045
(0020, 1040) Position Reference Indicator        LO: 'SN'
(0020, 9221)  Dimension Organization Sequence  1 item(s) ----
   (0020, 9164) Dimension Organization UID          UI: 1.3.6.1.4.1.43046.3.0.89502.1553284149.555763
```

Fig. 3.5: Example of some tags, elements, and values from a DICOM file in the LIDC-IDRI dataset [10].

3.3 Picture Archive and Communication System (PACS)

PACS are DICOM-driven medical systems with hardware and software built to facilitate digital imaging in clinical settings [171]. As mentioned earlier in this chapter, the major components of a diagnostic radiology workflow consist of imaging modalities, storage servers, and display stations. In a PACS system, these components are integrated by communication networks [83] and a viewing component supporting worklists for radiographers and radiologists. In addition, as illustrated in figure 3.6, PACS can further be connected with other information systems such as radiological information systems (RIS), an electronic medical records system designed to keep track of patient data and scan requests within radiology workflow [42].

The integration of PACS in radiology workflow has resolved many problems associated with traditional film-based diagnostic imaging. For example, Strickland [213] reported that radiology films were regularly unavailable when needed, with some hospitals reporting up to 20% of films being missing. With the PACS systems, patient studies can be safely stored in a server and accessed simultaneously by authorized physicians and radiological technologists from any authorized workstation [83].

The improved efficiency of diagnostic radiological workflow has allowed a higher volume of examinations to be requested, performed, and interpreted [161]. However, as discussed at the beginning of this chapter, the growth in the number of radiologists has not kept pace with the demand for their expertise [128]. In addition, the rising demand may also bring some technical challenges, such as image transmission problems, issues in data backup, and limited storage capacity issues [5].

3.4 Example of deep learning in radiological workflow

The shortage of radiologists and the increasing number of examinations requiring interpretation highlight the urgent need for new technologies. There's hope that deep learning-based applications integrated into clinical workflows can solve or at least alleviate these challenges. This is, together with the hope of improved diagnostic and prognostic accuracy based on individual characteristics of patients, the main reason for the immense interest in deep learning in radiology [149, 182, 230].

As we have discussed earlier in this thesis, it is clear that deep learning models show great potential in radiology, demonstrated by various applications [31, 58, 129], including those explored in this thesis. Moreover, collaborative setups including both humans and deep learning applications seem to yield even better performance and may reflect how radiologists work in clinical settings (e.g., the double reading process discussed in Chapter 1) [206, 247]. Figure 3.6 illustrates how deep learning applications can be applied at different stages of the diagnostic radiology workflow for patients with lung cancer.

Although deep learning applications show great potential in a wide range of medical imaging tasks, most of these studies are conducted in computer science laboratories using retrospective data [182, 230]. Retrospective data can be valuable for initial research. However, a significant problem with this approach is that it fails to evaluate how these tools will be used in clinical practice, potentially leading to inaccurate performance when encountering real-world data [4, 104].

Only a limited number of applications have been successfully deployed into existing radiology workflow [230]. For example, as discussed in Chapter 1, Leeuwen et al. [231] looked at 100 CE-marked AI products from 54 different vendors in 2020 and found that only 18 of these products had the potential to produce clinical impact. As of March 2023, over 200 CE marked deep learning applications are available on the market[2]. Yet, at the time of writing, only 44 of these applications have been approved under the new EU Medical Device Regulations (MDR) [167], for a total of 12 MRI- and 12 CT-based applications. To conform with MDR, AI applications must be thoroughly tested and validated in clinical studies, hopefully improving the quality of applications on the European market [155]. However, although MDR certification imposes stricter requirements for clinical applications, it does not necessarily imply clinical value (e.g., efficiency, costs, etc.).

In other words, to evaluate deep learning systems thoroughly, we need to integrate these applications into the real-world clinical workflow or as close as possible to the workflow where they will be utilized. As the field matures, more applications will likely be developed and evaluated in clinical practice.

We will discuss such opportunities and challenges further in Chapter 4.

[2] confer `https://grand-challenge.org/aiforradiology/` for a continuously updated list of deep learning-based products in radiology

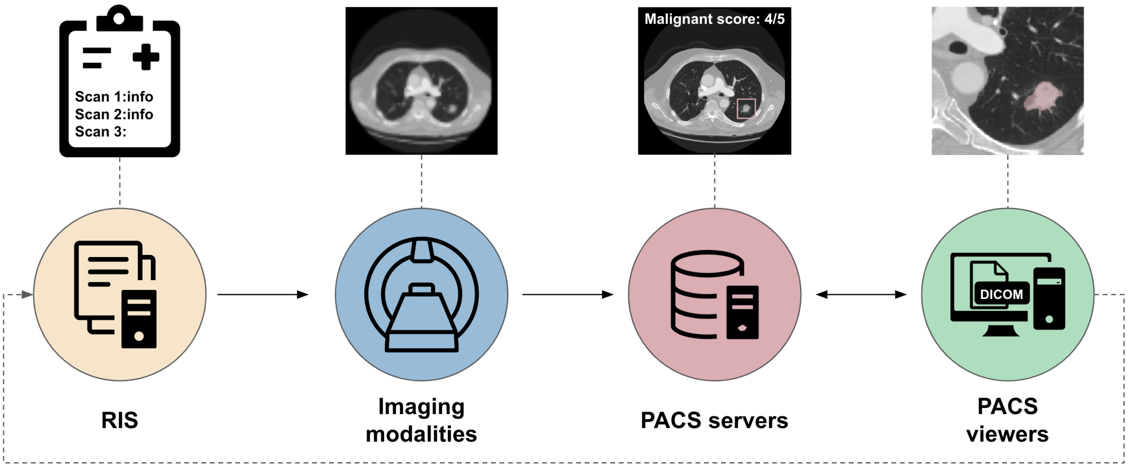


Fig. 3.6: Some potential use cases of deep learning applications for enhancing diagnostic radiology workflows. (1) Radiology information systems (RIS): Optimizing imaging appointment scheduling based on cancer-risk [84]. (2) Imaging modalities: Reducing radiation exposure and improving image quality through noise and artifact reduction methods in low-dose CT [251]. (3) PACS server: Employing applications to automatically detect potential cancer regions and calculate malignancy probabilities [9]. (4) PACS viewer: Utilizing segmentation tools [228] to efficiently evaluate the volume and appearance of regions of interest and further estimate lung cancer risk by integrating screening results with clinical data [84].

Code should be treated as a readable and executable form of literature, making the research process transparent and reproducible.

—ChatGPT (Jan 30 version)

CHAPTER 4

FROM PIXEL TO PATIENT: CONCLUSIONS, CONTRIBUTIONS, AND CONTINUATIONS

There is no doubt that deep learning has the potential to play an important role in the future of medicine [182, 225]. Deep learning models, especially convolutional neural networks (CNNs) and transformers, hold great potential for transforming medical imaging domains due to their remarkable performance on various computer vision tasks compared to traditional methods (see Chapter 2). Diagnostic radiology, in particular, has drawn special interest in the application of AI. However, as highlighted in Chapter 3, the field is still in the early phase: great promise but short on real-world clinical evaluation.

In discussing AI's role in medicine, it is crucial to differentiate between applying AI methodologies to medical data and what can properly be called "medical AI," namely the development, evaluation, and deployment of AI-based solutions in healthcare. For example, applying deep learning techniques to medical image data-based tasks can be relatively simple as long as you have access to data. This is exemplified by our one-click tutorials in fastMONAI [96] running MedMnist v2 [250] and Medical Segmentation Decathlon challenge [7] datasets. Such simple benchmark datasets play an essential role in the development of methods in various domains [221]. However, it is important to note that benchmark datasets often undergo extensive preprocessing and may not capture the complexity of real-world data [182], such as different imaging protocols, manufacturers, scanners, demographics, and many other factors. Furthermore, in practical settings, obtaining access to large amounts of labeled data, as seen in benchmark datasets, can be challenging due to the high cost of data annotation [242], which often requires domain experts. In other words, while these benchmark datasets can be valuable for initial research, they may not capture all the challenges in clinical settings.

Imaging-based medical AI is not really about images or the pixels within those images but about improving patient care. Therefore, researchers need to consider factors beyond model performance on benchmark datasets to develop medical AI tools and systems that can be valuable in clinical settings for both clinicians and patients. These factors include problem understanding, data understanding, and integration of AI systems with existing workflows (see figure 4.1). Each of these components consists of many practical, technical, and ethical challenges [30, 73, 148]. Interdisciplinary collaboration is critical to addressing these challenges. In this thesis, I have had the opportunity to ensure greater clinical relevance through close collaboration with radiologists and healthcare IT infrastructure experts at Mohn Medical Imaging and Visualization Center (MMIV) at Haukeland University Hospital (HUH).

This chapter will draw some conclusions from earlier chapters' considerations and discuss important challenges and opportunities in the field. Along the way, we will look at the contributions from my thesis and point to possible future work.

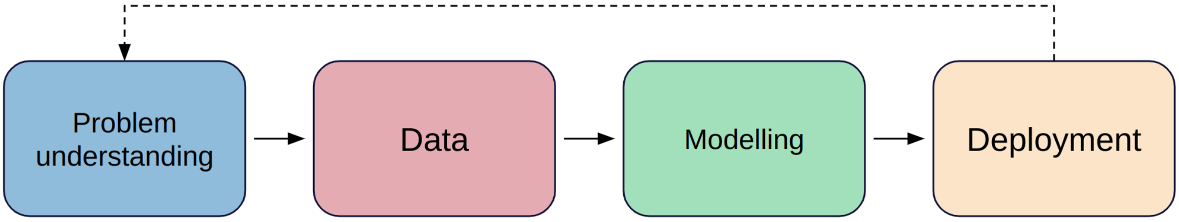


Fig. 4.1: A diagram highlighting the different stages in a deep learning lifecycle: **(1)** Problem understanding: identifying the intended use and clinical role of the deep learning application, defining data sources, and establishing guidelines for annotation; **(2)** Data preparation: data collection, labeling, and data preprocessing; **(3)** Model development: selection of appropriate model, training strategies, and model evaluation; **(4)** Deployment: integrating the trained model into existing infrastructure and monitoring its performance. These stages form an iterative process, ensuring that the models remain up-to-date. See Table 4.1 for a thorough description.

4.1 Data-centric AI: Deep learning = data + models

A supervised deep learning system's performance is determined and limited by the models, data, and corresponding labels used to train and evaluate them. However, as highlighted in Chapter 2, over the past decade, the research focus in the field has primarily revolved around model-centric developments. New architectures and training strategies have constantly been proposed and reported to outperform the prior state-of-the-art on a wide range of large benchmark datasets (e.g., the ImageNet dataset, Microsoft COCO [121], CIFAR-100, etc.). There is no doubt that this model-centric approach has driven the field forward. This progress has led to the introduction of new architectures such as ResNet [72] and U-Net [191], and more recently, Swin Transformer [125] and ConvNext [126]. Additionally, several training strategies have also emerged, including data augmentation techniques like CutMix [253] and MixUp [257], Batch Normalization [90], Dropout [211], and many others.

However, as the field matures, we observe a shift in focus toward data-centric development, driven by various important concerns. These include issues like bias and fairness [3, 47], privacy [200], and label quality [43]. As reported in [157], many commonly used benchmark datasets contain label errors in both training and validation sets due to the inherent construction process, which often involves some degree of automatic labeling or crowd-sourcing.

Some studies have shown that label noise is not a major problem in the training set, as deep learning methods are believed to be inherently robust to label noise when a large amount of data is available [131, 189, 215]. However, as highlighted by Rolnick et al. [189], clean labels lead to better performance than noisy labels, given the same quantity of training data.

The presence of label noise in validation data may lead to unreliable model evaluation and inaccurate estimation of a model's performance. For example, multiple studies in recent years have highlighted label issues in the ImageNet validation set [21, 185, 227]. According to Northcutt et al. [157], about 6% of the validation set is incorrectly labeled, which makes models able to perform with an error rate of less than 6% inherently suspicious. In figure 4.2, we present examples of label noise from the ImageNet validation set. It is essential to address the issue of incorrect labeling in validation and test sets, as it could potentially lead to incorrect conclusions about model performance in the real world.

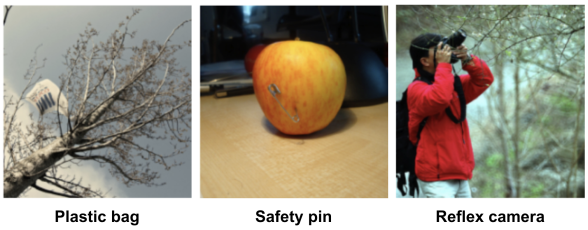


Fig. 4.2: Some examples from the ImageNet validation set with their corresponding labels that our ResNet34 models were uncertain about. As we can see, these images should ideally have multiple labels assigned to them. We have provided source code to find similar examples on GitHub: https://github.com/skaliy/thesis_supplementary_materials.

We have had to deal with this also in our own work. In our brain extraction study [95], label noise was observed in the training/validation set, as we used labels automatically generated by Brain Extraction Tool (BET) [208] of the FMRIB Software Library (FSL) [245].

To cope with this issue, we trained a model on the entire training set (training and validation) for a few epochs. We then manually inspected the images that the model was most uncertain about. By applying this approach, we ended up removing 14 images from the training set before training our final model. Finally, when we evaluated the performance of our model on the test data, we found 12 additional images that were clear FSL failures, not prediction failures, as confirmed by visual inspection.

In medical imaging diagnostics, the stakes are exceptionally high, as model performance could directly impact patient care and clinical decision-making. Unfortunately, as reported in [100, 149, 154], developing high-quality labeled data in the medical field can be challenging and expensive due to factors such as difficulty in acquiring data, the need for manual annotation by radiologists, and the lack of efficient labeling frameworks. Medical image datasets typically end up being very small, particularly 3D medical datasets, ranging from a few hundred to a few thousand images [2, 7, 10, 12, 13, 143, 198][1].

Moreover, there is often significant inter- and intraoperator variability between radiologists, which can lead to inconsistencies in the labeling process [142, 143, 238]. The issue is demonstrated in figure 4.3, using the bottle segmentation example from Chapter 2. Even such a simple task can expose label discrepancies between raters, emphasizing the complexity of the real world. In the context of medical image diagnostics tasks, such as brain tumor segmentation (figure 3.3), the process is prone

[1] A list of public datasets used in this thesis can be found at https://github.com/skaliy/thesis_supplementary_materials

to subjectivity [142, 143, 238]. In cases where multiple raters provide labels, simple voting methods, such as majority voting or more advanced weighting methods like STAPLE (Simultaneous Truth and Performance Level Estimation) [236] can be employed to generate consensus-based ground-truth. However, as highlighted by Warfield et al. [236], issues caused by rater-specific bias may remain. Consequently, prior to initiating the manual annotation process for a new dataset for deep learning purposes, establishing well-defined and consistent guidelines can help address potential challenges and improve the quality of the labels.

A number of important areas within medical imaging and imaging diagnostics have not yet experienced the impact of the latest developments in deep learning. This is often because most researchers focus on areas where large amounts of annotated data are already available or where the annotation process is reasonably easy to carry out, which does not necessarily capture the areas of the greatest importance and surely does not capture them all [233].

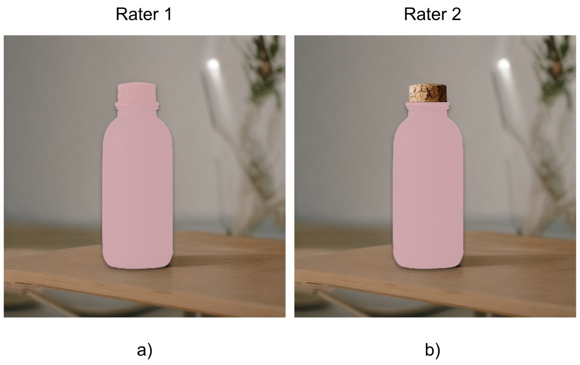


Fig. 4.3: Illustration demonstrating inter-rater variability in the segmentation of the same bottle by two different raters when conducted without carefully designed guidelines.

Willemink et al. [242] emphasize that, in research settings, a limited amount of data could be adequate for training models. As showcased in our clinical studies on cervical cancer [75] and spine segmentation [94], the number of examples needed to learn a new task is highly dependent on the complexity of the task, with the first one requiring many more labeled instances than the latter.

Considering the high costs associated with label processing, there are several human-in-the-loop methods and strategies to reduce annotation time and enlarge training datasets, including semi-automatic labeling and active learning.

Semi-automatic labeling can involve using an initially trained model to predict labels for a larger, unlabeled dataset. Experts can further examine and adjust these suggested labels as needed and iteratively update the model by incorporating them into the training set.

Once some manually or semi-automatically labeled examples exist, active learning can be applied to further train the model. Active learning is a set of iterative techniques that focus on identifying and prioritizing the most informative samples from unlabeled datasets for human annotators to label [28, 204]. Several query methods have been developed to efficiently identify these informative samples. Uncertainty sampling is one of the most popular strategies, which aims to select samples that the current model is most uncertain about to maximize model improvement [16, 92, 235].

Once a dataset has been created and the model has been fully trained, sharing the data and trained weights within the research community can be highly valuable for other researchers and developers. However, when it comes to medical data, privacy, ethical, and legal concerns need to be taken into account, making data sharing challenging [18, 232, 246].

A solution to overcome this obstacle is to share the code and trained model weights, along with comprehensive documentation about the dataset. As emphasized by Aggarwal [4], the lack of external validation in many studies is a significant concern. Sharing code and trained weights can promote external validation and transparency in research, enabling other researchers to evaluate the model's performance using their own data.

Moreover, transfer learning enables other researchers to build on the knowledge acquired from the pre-trained model by fine-tuning it for their specific tasks using a smaller dataset [237, 252, 261]. This approach minimizes the number of labeled examples required in the target domain and could also improve learning performance. This may bridge the gap between researchers with access to vast computational resources and those who may be limited by hardware or sample size.

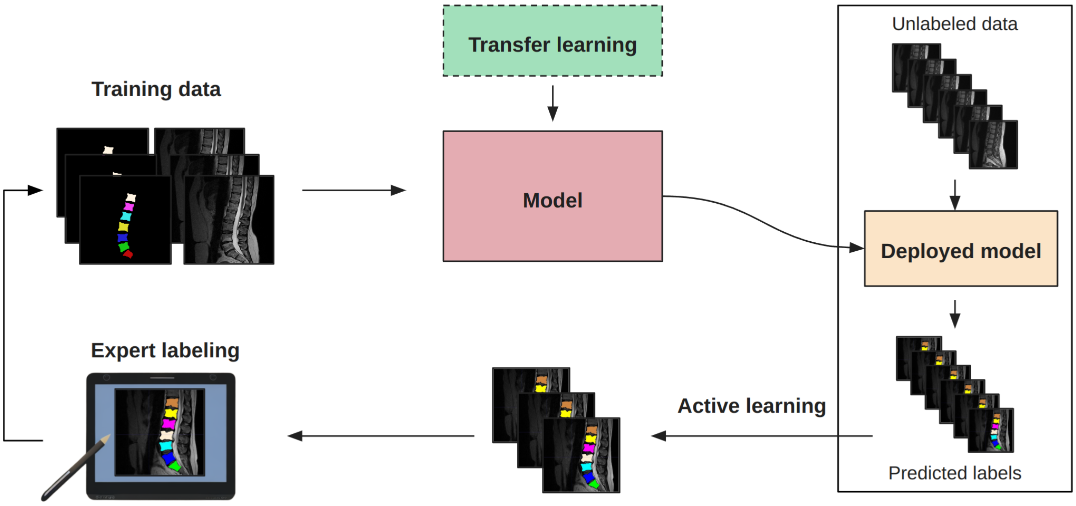


Fig. 4.4: Human-in-the-loop pipeline for deep learning model development. The process begins with training a model using available data (and utilizing transfer learning with a pre-trained model, if available). After training, the model is deployed to generate predictions for new, unlabeled data. An active learning approach can be applied here to identify the most informative samples, which an expert then refines. These data are then used to retrain the model, enabling continuous improvement of the model's performance through iterative expert feedback.

4.2 State-of-the-art-mania

As we have seen in Chapter 2, there is a constant influx of new models claiming to outperform the previous state-of-the-art in various benchmarks. However, the field of deep learning suffers from what are often unfair comparisons between existing models and methods and new ones. The research by Liu et al. [126] on ConvNext emphasizes the importance of fair comparison. Their work demonstrated that vision transformers believed to outperform CNNs failed to do so when evaluated under fair conditions (i.e., applying the same training techniques and design methods).

The publication trend (as shown in table 2.2) is to report better results with larger and more scalable models. There are some concerns with this that we want to highlight in this section.

To add novelty to their research, scientists may incorporate unnecessary complexity into their methods [233]. However, this complexity may not necessarily enhance model performance, but instead, contributes to technical debt, making systems more challenging to deploy and maintain [203][2].

This observation is supported by the success of the nnU-Net ("no-new-Net") framework from 2021 [91], which uses a model architecture from 2015 (discussed in Chapter 2). The nnU-Net framework achieved leading performance in a wide range of medical image segmentation tasks and is widely used in medical imaging research.

Focusing on larger and more scalable models results in an increased demand for computational infrastructure, which is rarely reported in the literature. Although the source code could be made available, there is no guarantee to reproduce the results due to the cost of reproducing the research. Some of these large computer vision models cost millions of dollars to train and use immense amounts of resources [50, 222]. Moreover, the rising cost associated with these models may restrict cutting-edge technology development and availability to a few organizations (e.g., Microsoft, Google, OpenAI, Meta, etc.). This trend is reflected in the large-scale models published today [25, 105, 163, 223]. However, there are exceptions. For example, the Computer Vision Group and Learning research group at the Ludwig Maximilian University of Munich developed Stable Diffusion [190] as an alternative to DALL-E 2 [183], making their trained model and source code openly available. Consequently, several new ideas, improvements, and insights have emerged from the community [24, 26, 34], demonstrating the positive impact of open science.

The research community must recognize that the push for larger and scalable models is highly resource-intensive, and deploying these models in the real world may pose significant challenges. The severity of this problem is well illustrated by the case of OpenAI's decision not to correct a mistake implemented in GPT-3 due to the vast resource requirements [222]. This is particularly relevant in deploying models in the field of medicine, as model performance could directly impact patient care and clinical decision-making.

4.3 From model to deployment: Navigating through the minefield

As we discussed in Chapter 3, although deep learning models developed in computer science laboratories can be accurate on retrospective data, a major problem with this approach is that it fails to evaluate how these tools will be used, potentially leading to suboptimal performance when applied to real-world clinical data.

Researchers should also consider other factors beyond model performance, such as speed, efficiency, and cost reduction [4]. Moreover, there are various ethical concerns that need to be addressed when implementing deep learning models in radiology, including fairness, accountability, and transparency [152].

To thoroughly evaluate deep learning models, it is crucial to integrate them into real-world workflows and study the aforementioned factors in prospective studies [104].

[2] The fourth rule of machine learning is *"Keep the first model simple and get the infrastructure right"* [262].

This allows researchers and clinicians to identify potential pitfalls and opportunities for improvement before deploying such systems in real clinical settings.

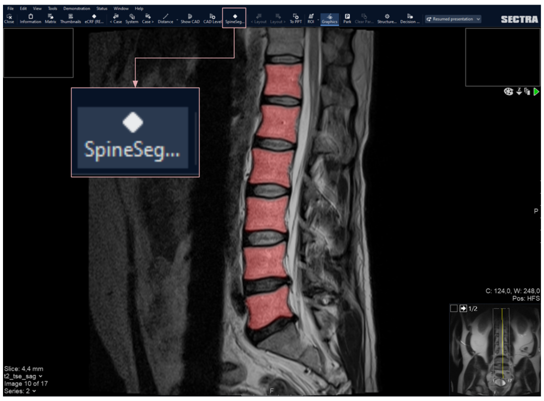


Fig. 4.5: Our spine segmentation application deployed in the research PACS system. Given a compatible spinal MRI recording, a single button click runs our deep learning-based segmentation model, and the results are displayed to the user inside the PACS workstation.

Despite its importance, deploying deep learning models into radiology workflow continues to be a challenge [149, 230]. One contributing factor is the technical obstacles encountered when attempting to seamlessly integrate these models into the existing infrastructure, such as containerizing models and orchestrating with other systems (e.g., Docker [144] or Kubernetes [29]) and reading and storing DICOMs correctly in PACS.

To address these integration challenges, we have been working closely with researchers from the Workflow-Integrated Machine Learning (WIML) [147] project at MMIV, aiming to implement deep learning projects in the Western Norway Regional Health Authorities (Helse Vest RHF), a network of four major hospitals and about 30 healthcare institutions. A significant component in the WIML infrastructure is the research PACS.

As the name implies, the research PACS is explicitly designed to house research project data imported as pseudonymized scans from the clinical PACS[3]. The data in research PACS are stored as DICOM with an enhanced DICOM data model, including the project name and data collection visit information.

The first deep learning-based workflow in the system was created by Digernes and Ditlev-Simonsen [51], applying an early version of fastMONAI [96]. They used the BraTS brain tumor segmentation dataset [12, 13, 143] to construct a tumor segmentation model to assess the workflow's obstacles and possibilities. We have built upon this work and added functionalities for running models automatically using APIs from research PACS for projects within Western Norway Regional Health Authorities. Figure 4.5 displays a screenshot of our spine segmentation model [94] applied to an unseen subject from the clinical trial AIM (Antibiotics In Modic Changes) [212].

As discussed in the study, we observed some performance variability among subjects. A natural next step would involve assessing the model's performance on a larger cohort within this system and investigating the causes of this variability (e.g., differences in scanner settings, anatomical variation, and other factors). To label a more extensive dataset in future studies, we plan to implement an active learning setup in research PACS that identifies and prioritizes the most informative samples for human experts to label (see Section 4.1).

Moreover, in future work, we aim to compare our current solution with MONAI Deploy [33], designed for deploying deep learning-based applications into clinical

[3] The hospitals in Helse Vest RHF utilize Sectra IDS7 as their PACS solution.

workflows [4].

4.4 Publishing, open science, and reproducibility

The number of AI publications over the past decade has been rising rapidly, resembling an exponential growth curve [256]. However, machine learning-based science is faced with major reproducibility issues. In [99], the authors analyzed 20 review papers across 17 different research fields, focusing on errors when applying machine learning, and found that a total of 329 research papers were not fully reproducible due to how machine learning was implemented in their work. In some cases, this resulted in overoptimistic performance claims. Huston [88] reported a similar concern when attempting to compare their model against what was reported to be state-of-the-art. Since the source code for the project they wanted to compare with was not published, the researchers had to write their own code based on the description in the paper. However, even after two months of work, they could not get close to the reported results[5].

This example shows the importance of transparency in machine learning research. When source code and data are not openly shared, it becomes challenging for other researchers to verify the validity of the reported results or build upon existing work to make progress in the field. The lack of transparency can lead to wasted time, effort, and resources as researchers attempt to replicate findings based on limited information. Moreover, the lack of transparency hinders the scientific community's ability to identify potential flaws, biases, and inconsistencies in the reported results [99].

Although sharing source code and data is widely acknowledged as a crucial component of scientific research, over the years, many researchers in the deep learning field seem to have prioritized getting their findings published and neglected to share their code, data, and documentation that is necessary for others to reproduce their work. This is perhaps especially true for many industry researchers [41, 67, 122, 151].

For example, three years ago, a group of researchers from Google Health demonstrated the potential use of deep learning models for breast cancer screening [141]. However, despite the promising results, an investigation by Haibe-Kains et al. [67] revealed that the study was missing crucial details about the data processing, model development, and training pipelines, making it difficult to reproduce their work. As highlighted in their report, this restricts the scientific impact and hinders other researchers from building upon it in future studies.

In response to this critique, Google Health researchers acknowledged the importance of transparency and reproducibility in scientific research [139]. As a result, they provided additional methodological details in an addendum to the original article [140]. A noteworthy point in their response was that most of their work builds on open-source software implementations.

[4] This video gives an overview and demonstration of MONAI Deploy, recorded during the MONAI Bootcamp 2023: `https://youtu.be/mpVEiNW9qtw`. In the video, the presenter discusses plans to create short tutorials for the various elements of the deployment process, aiming to help users better understand the framework

[5] This is something I, and likely most other machine learning researchers [62], have experienced many times.

In Chapter 3, we discussed the rapid emergence of commercial AI applications in diagnostic radiology. Despite this growth, only a few applications have till now shown clinical value. Moreover, to the author's knowledge, most of these applications are closed source and provide limited information on product specifications (some refer to peer-reviewed papers on performance, but these are short on implementation details). For instance, in one MDR-approved breast cancer detection application, the vendor describes the application as suitable for *women at any age* but fails to provide further details.

Understandably, vendors might be unable or unwilling to disclose all the information about their applications due to various reasonable concerns, including competition in the market. However, as emphasized by Varoquaux [233], models in medicine should document their limitations and the choices made during their training process. One way to provide more information could be to adopt the concept of *model cards* [146]. Model cards provide short documentation of a model's strengths and limitations (e.g., performance, potential biases, etc.), allowing researchers and developers to understand the application better. Note that the proposed EU Artificial Intelligence Act is relevant to these issues. However, the details remain ambiguous [56, 57].

Such transparency would build trust in the application's validity, potentially accelerating the adoption in clinical settings and ultimately improving patient outcomes.

This is, of course, also true for research-oriented work. For example, Rajpurkar et al. [182] emphasize that standards for transparency in reporting and validation are needed to build trust in deep learning-based research. Luckily, numerous platforms exist to make deep learning research more transparent and reproducible, including code-sharing platforms and tools for sharing trained models.

In recent years, many AI conferences and journals have encouraged authors to share code and details about the data used to construct their methods by implementing reproducibility checklists [73, 148, 173]. In table 4.1, we re-print a valuable guide for the development of deep learning models for medical image analysis (Checklist for Artificial Intelligence in Medical Imaging, CLAIM), developed by Mongan et al. [148].

Table 4.1: Checklist for Artificial Intelligence in Medical Imaging (CLAIM) Mongan J, Moy L, Kahn C E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers [148]. Radiology: Artificial Intelligence 2020. Published online on March 25, 2020. DOI: 10.1148/ryai.2020200029. Table reproduced as originally published with permission from ©Radiological Society of North America.

Section/Topic	**No.**	**Item**
TITLE or ABSTRACT		
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)
ABSTRACT		

Continued on next page

Section/Topic	**No.**	**Item**
	2	Structured summary of study design, methods, results, and conclusions
INTRODUCTION		
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach
	4	Study objectives and hypotheses
METHODS		
Study Design	5	Prospective or retrospective study
	6	Study goal, such as model creation, exploratory study, feasibility study, noninferiority trial
Data	7	Data sources
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)
	9	Data preprocessing steps
	10	Selection of data subsets, if applicable
	11	Definitions of data elements, with references to common data elements
	12	De-identification methods
	13	How missing data were handled
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication
	15	Rationale for choosing the reference standard (if alternatives exist)
	16	Source of ground truth annotations; qualifications and preparation of annotators
	17	Annotation tools
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies
Data Partitions	19	Intended sample size and how it was determined

Continued on next page

Section/Topic	No.	Item
	20	How data were assigned to partitions; specify proportions
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections
	23	Software libraries, frameworks, and packages
	24	Initialization of model parameters (e.g., randomization, transfer learning)
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained
	26	Method of selecting the final model
	27	Ensembling techniques, if applicable
Evaluation	28	Metrics of model performance
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)
	30	Robustness or sensitivity analysis
	31	Methods for explainability or interpretability (e.g., saliency maps) and how they were validated
	32	Validation or testing on external data
RESULTS		
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion
	34	Demographic and clinical characteristics of cases in each partition
Model performance	35	Performance metrics for optimal model(s) on all data partitions
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	37	Failure analysis of incorrectly classified cases
DISCUSSION		

Continued on next page

Section/Topic	No.	Item
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability
	39	Implications for practice, including the intended use and/or clinical role
OTHER INFORMATION		
	40	Registration number and name of registry
	41	Where the full study protocol can be accessed
	42	Sources of funding and other support; role of funders

A report [173] from the NeurIPS 2019 reproducibility program revealed that when NeurIPS introduced a reproducibility checklist, the number of papers with code submitted increased from <50% to 75%. Moreover, many reviewers explored the code when evaluating submissions.

While sharing source code is crucial for reproducibility, it is not enough. To reproduce and validate an experiment, you need the exact data used to train and evaluate the reported model. Unfortunately, as discussed earlier in this thesis, this is not always possible in domains such as medicine for various reasons (e.g., patient privacy, laws, etc.). To address this limitation, the above-mentioned report [173] suggests that researchers should provide complementary empirical results on an open-source dataset alongside the results from any confidential dataset.

In this thesis, we have strived to take steps toward making the included publications reproducible and transparent through sharing code, applying open-source data for evaluation, and describing datasets in detail in cases where we could not share or find open-source datasets[6]. In our spine segmentation study [94], we share not only the source code but also the trained weights and a model card, enabling other researchers to understand the model and its limitations better, validate the reported results and generate segmentation masks for their own data.

Furthermore, our fastMONAI article is provided as a *computational essay* [52, 160, 243]. In short, an essay that combines text and illustrations with executable code. Computational essays enable readers to explore and validate ideas by rerunning computations themselves. We use Jupyter Notebook [106] (described in Chapter 1) to achieve this integration. Furthermore, the paper is made available on Google Colab [23], enabling users to run the notebook on the cloud using free GPU instances and effortlessly install the required libraries with one click using any computing device with a web browser (including smartphones).

[6] To the best of the author's knowledge, there are no open-source data available with similar sequences for cervical cancer as used in our study [75]. We looked for prostate cancer datasets with similar MRI sequences, but none with labels were found at the time of conducting the study. In future research, we plan to examine the Prostate158 [2] and PI-CAI (Prostate Imaging: Cancer AI) [198] datasets (published after our study) for a potential transfer learning-based study.

Using Jupyter Notebook and cloud computing services for writing articles gives readers an interactive and dynamic platform for engaging with the content, gaining a deeper understanding of the concepts presented in the article, and exploring related ideas. In future work, we plan to use this setup for other research articles in cases where it is feasible and useful and encourage other researchers to do the same.

CHAPTER 5

SUMMARY OF PAPERS

This chapter contains a summary of the articles produced during the development of this thesis. The main contribution of the thesis is paper A, which introduces our library, fastMONAI. In all our other work (B, C, D, E), we used fastMONAI and our precursors that culminated in fastMONAI. In addition, the author of this thesis has published a paper related to remote sensing [97]. This work is not included in the thesis, but it serves as a demonstration of the generic nature of deep learning methods.

Figure 5.1 provides an overview of all the articles included in this thesis.

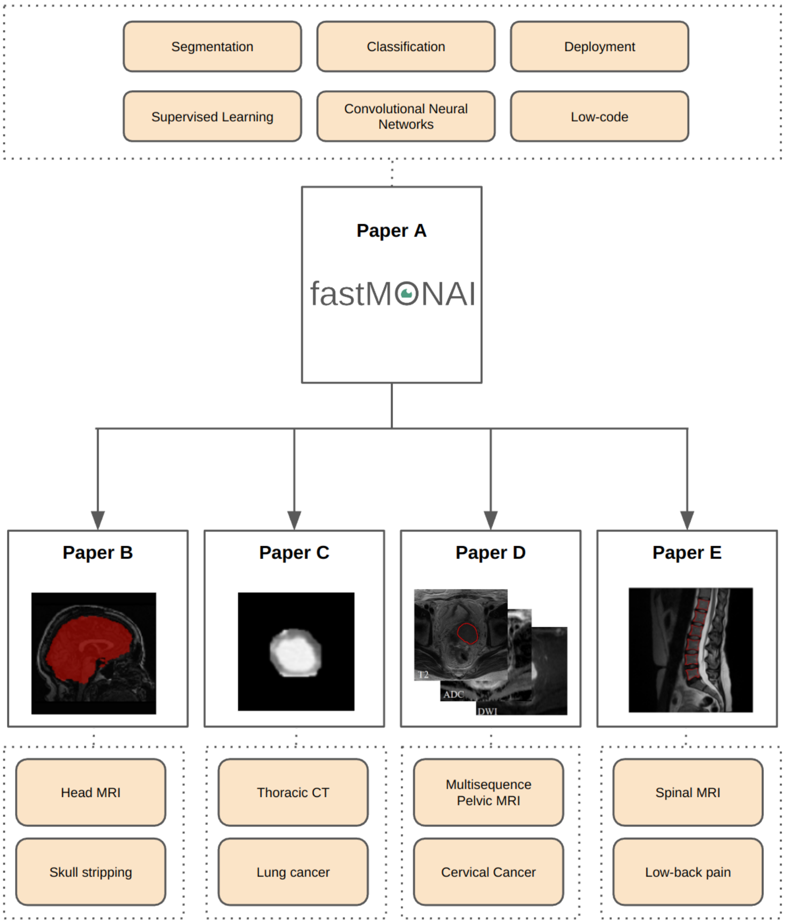


Fig. 5.1: The illustration gives an overview of the thesis. Paper A presents our fastMONAI library. The library was constructed as a result of the research projects reported in Papers B to E and forms a core methodological component of these projects.

5.1 Paper A: fastMONAI: a low-code deep learning library for medical image analysis

In paper A [96], we present fastMONAI, a low-code Python-based open-source deep learning library built on top of fastai [80], MONAI [33], and TorchIO [170]. We created the library to simplify the use of state-of-the-art deep learning techniques in 3D medical image analysis for solving classification, regression, and segmentation tasks. fastMONAI provides users with functionalities to step through data loading, preprocessing, training, and result interpretations. The entire library is developed using nbdev [81], a tool for exploratory programming that allows you to write test and document a Python library in Jupyter Notebooks [106].

The paper was automatically generated from a Jupyter Notebook available in the fastMONAI GitHub repo: `https://github.com/MMIV-ML/fastMONAI`. By using the notebook version of the paper, the reader can step through the paper's content and reproduce all the text, computations, figures, and results.

5.2 Paper B: 2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using fastai

Skull stripping in brain imaging is the removal of the parts of images corresponding to non-brain tissue. Fast and accurate skull stripping is crucial for numerous medical brain imaging applications, e.g., registration, segmentation, and feature extraction, as it eases subsequent image processing steps. In paper B [95], we propose and compare two novel skull stripping methods based on 2D and 3D convolutional neural networks trained on a large, heterogeneous collection of 2777 clinical 3D T1-weighted MRI images from 1681 healthy subjects. We investigated the performance of the models by testing them on 927 images from 324 subjects set aside from our collection of data, in addition to images from an independent, large brain imaging study: the IXI dataset (n = 556). Our models achieved mean Dice scores higher than 0.978 and Jaccard indices higher than 0.957 on all test sets, making predictions on new unseen brain MR images in approximately 1.4s for the 3D model and 12.4s for the 2D model. In addition, a preliminary exploration of the model's robustness to variation in the input data showed favorable results when compared to a traditional, well-established skull stripping method. With further research to increase the models' robustness, such accurate and fast skull stripping methods can potentially form a valuable component of brain MRI analysis pipelines. A list of all the data sources used in our study is available on GitHub: `https://github.com/MMIV-ML/Skull-stripping-NIK2020`[1].

5.3 Paper C: Pulmonary nodule classification in lung cancer from 3D thoracic CT scans using fastai and MONAI

In paper C [98], we construct a convolutional neural network to classify pulmonary nodules as malignant or benign in the context of lung cancer. To build and train

[1] A tutorial for constructing a skull-stripping model is provided in paper A

our model, we use our novel extension of the fastai deep learning framework to 3D medical imaging tasks combined with the MONAI deep learning library. We train and evaluate the model using an extensive, openly available annotated thoracic CT scan data set. Our model achieves a nodule classification accuracy of 92.4% and a ROC AUC of 97% when compared to a "ground truth" based on multiple human raters' subjective assessment of malignancy. We further evaluate our approach by predicting patient-level cancer diagnoses, achieving a test set accuracy of 75%. This is higher than the 70% obtained by aggregating the human raters' assessments. Finally, class activation maps are applied to investigate the features used by our classifier, enabling a rudimentary level of explainability for what is otherwise close to "black box" predictions. As the classification of structures in chest CT scans is useful across various diagnostic and prognostic tasks in radiology, our approach has broad applicability. We aimed to construct a fully reproducible system that can be compared to newly proposed methods and easily be adapted and extended; the complete source code of our work is available at https://github.com/MMIV-ML/Lung-CT-fastai-2020.

5.4 Paper D: Fully automatic whole-volume tumor segmentation in cervical cancer

Uterine cervical cancer (CC) is the most common gynecologic malignancy worldwide. Whole-volume radiomic profiling from pelvic MRI may yield prognostic markers for tailoring treatment in cervical cancer. However, radiomic profiling relies on manual tumor segmentation, which is unfeasible in the clinic. In paper D [75], we present a fully automatic method for the 3D segmentation of primary cervical cancer lesions using state-of-the-art deep learning techniques.

In 131 cervical cancer patients, the primary tumor was manually segmented on T2-weighted MRI by two radiologists (R1, R2). The patients were split into a training/validation (n = 105) and a test- (n = 26) cohort. The deep learning model's segmentation performance, when compared with R1/R2, was lower in terms of both median Dice coefficients (DSCs) (DL-R1 = 0.60, DL-R2 = 0.58, R1-R2 = 0.78) and median Hausdorff distances (DL-R1 = 29.2 mm; DL-R2 = 30.2 mm, R1-R2 = 14.6 mm) in the test cohort.

Although the achieved segmentation performance of the trained deep learning model is slightly lower than that for radiologists, this study demonstrates its potential to enable automated estimation of tumor size and primary cervical cancer tumor segmentation. The source code is available at https://github.com/MMIV-ML/cervical-cancer-segmentation-2022.

5.5 Paper E: Multi-Center CNN-based spine segmentation from T2w MRI using small amounts of data

Segmentation of the spinal tissues on MRI is the basis for quantitative analyses, but time-consuming if done manually. In paper E [94], we construct a pipeline for automatic vertebrae segmentation from T2w MRI scans, assessing performance and generalizability by external validation. Our study used 15 scans from one site

(Haukeland University Hospital, HUH) and 10 scans from another (Sahlgrenska University Hospital, SUH). MRI experts manually delineated the vertebral bodies Th12-L5 on all the HUH data and a subset of six scans from SUH. We trained multiple convolutional neural networks, assessing the performance in an experimental design tailored to small-data contexts and also on external data. Our best model achieved a mean Dice score of 0.899. This is comparable to results in the literature, but our system required much less training data. Furthermore, the trained model can form a component in an active learning setup to lower the time needed for manual delineation. The source code is available at `https://github.com/MMIV-ML/fastMONAI/tree/master/research`.

BIBLIOGRAPHY

[1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. Multimodal Biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022. 1.1

[2] L. C. Adams, M. R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S. M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, et al. Prostate158-An Expert-annotated 3T MRI Dataset and Algorithm for Prostate Cancer Detection. *Computers in Biology and Medicine*, 148:105817, 2022. 4.1, 6

[3] C. C. Aggarwal. Neural Networks and Deep Learning. *Springer*, 10(978):3, 2018. 2.4, 2.4, 4.1

[4] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi. Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-analysis. *NPJ Digital Medicine*, 4(1):65, 2021. 1.1, 3.4, 4.1, 4.3

[5] M. Alhajeri and S. G. S. Shah. Limitations in and Solutions for Improving the Functionality of Picture Archiving and Communication System: an Exploratory Study of PACS Professionals' Perspectives. *Journal of Digital Imaging*, 32(1):54–67, 2019. 3.3

[6] E. Andersen. Imagedata: A Python Library to Handle Medical Image Data in NumPy Array Subclass Series. *Journal of Open Source Software*, 7(73):4133, 2022. 3.2

[7] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The Medical Segmentation Decathlon. *Nature communications*, 13(1):4128, 2022. 3.3, 4, 4.1

[8] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On Instabilities of Deep Learning in Image Reconstruction and the Potential Costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. 3

[9] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. End-to-end Lung Cancer Screening With Three-dimensional Deep Learning on Low-dose Chest Computed Tomography. *Nature medicine*, 25(6):954–961, 2019. 3.6

[10] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical physics*, 38(2):915–931, 2011. 3.1.1, 3.2, 3.2, 3.5, 4.1

[11] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann. Data Excellence for AI: Why Should You Care? *Interactions*, 29(2):66–69, 2022. 1.1

[12] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017. 3.3, 4.1, 4.3

[13] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv preprint arXiv:1811.02629*, 2018. 3.3, 4.1, 4.3

[14] S. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee. Impact of Fully Connected Layers on Performance of Convolutional Neural Networks for Image Classification. *Neurocomputing*, 378:112–119, 2020. 2.4

[15] M. F. Bellolio, H. C. Heien, L. R. Sangaralingham, M. M. Jeffery, R. L. Campbell, D. Cabrera, N. D. Shah, and E. P. Hess. Increased computed tomography utilization in the emergency department and its association with hospital admission. *Western Journal of Emergency Medicine*, 18(5):835, 2017. 1.1, 3

[16] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 4.1

[17] Y. Bengio, P. Simard, and P. Frasconi. Learning Long-term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. 2.2

[18] H. B. Bentzen, R. Castro, R. Fears, G. Griffin, V. Ter Meulen, and G. Ursin. Remove obstacles to sharing health data with researchers outside of the European Union. *Nature Medicine*, 27(8):1329–1333, 2021. 1.1, 4.1

[19] E. Bercovich and M. C. Javitt. Medical Imaging: From Roentgen to the Digital Revolution, and Beyond. *Rambam Maimonides Medical Journal*, 9(4), 2018. 3, 3.1

[20] M. A. Bernstein, K. F. King, and X. J. Zhou. *Handbook of MRI Pulse Sequences*. Elsevier, 2004. 3.1.2

[21] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are We Done With ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. 4.1

[22] W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212, 1997. 3.2

[23] E. Bisong et al. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 2019. 4.4

[24] D. Bolya and J. Hoffman. Token Merging for Fast Stable Diffusion. *arXiv preprint arXiv:2303.17604*, 2023. 4.2

[25] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021. 1.1, 4.2

[26] T. Brooks, A. Holynski, and A. A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800*, 2022. 4.2

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2.2

[28] S. Budd, E. C. Robinson, and B. Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021. 4.1

[29] B. Burns, J. Beda, K. Hightower, and L. Evenson. *Kubernetes: up and running*. " O'Reilly Media, Inc.", 2022. 4.3

[30] F. Cabitza and A. Campagner. The Need to Separate the Wheat From the Chaff in Medical Informatics: Introducing a Comprehensive Checklist for the (Self)-assessment of Medical AI Studies. *International Journal of Medical Informatics*, 153:104510, 2021. 4

[31] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy. Deep Learning for chest X-ray Analysis: A Survey. *Medical Image Analysis*, 72:102125, 2021. 1.1, 3.4

[32] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023. 2.2

[33] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. MONAI: An Open-source Framework for Deep Learning in Healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 1.1, 4.3, 5.1

[34] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting Training Data from Diffusion Models. *arXiv preprint arXiv:2301.13188*, 2023. 4.2

[35] J. Chai, H. Zeng, A. Li, and E. W. Ngai. Deep Learning in Computer Vision: A Critical Review of Emerging Techniques and Application Scenarios. *Machine Learning with Applications*, 6:100134, 2021. 2.1

[36] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative Pretraining from Pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 2.1, 2.5

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2.1

[38] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, page 102444, 2022. 2.1

[39] R. H. Choplin, J. Boehme 2nd, and C. D. Maynard. Picture Archiving and Communication Systems: An Overview. *Radiographics*, 12(1):127–129, 1992. 3

[40] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2.2

[41] J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi. The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—opportunities, Challenges, and Recommendations. *Ai & Society*, pages 1–25, 2021. 4.4

[42] J. Crabbe, C. Frank, and W. Nye. Improving Report Turnaround Time: An Integrated Method Using Data From a Radiology Information System. *AJR. American Journal of Roentgenology*, 163(6):1503–1507, 1994. 3.3

[43] K. Crawford and T. Paglen. Excavating AI: The Politics of Images in Machine Learning Training Sets. *Ai & Society*, 36(4):1105–1116, 2021. 4.1

[44] L. Curiel, R. Chopra, and K. Hynynen. Progress in Multimodality Imaging: Truly Simultaneous Ultrasound and Magnetic Resonance Imaging. *IEEE Transactions on Medical Imaging*, 26(12):1740–1746, 2007. 3.1

[45] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Advances in Neural Information Processing Systems*, 29, 2016. 2.1

[46] L. S. Davis. A Survey of Edge Detection Techniques. *Computer Graphics and Image Processing*, 4(3):248–270, 1975. 2.1

[47] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten. Does Object Recognition Work for Everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019. 2.1, 4.1

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2.1, 2.4, 2.1, 2.2

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2.1, 2.2

[50] P. Dhar. The Carbon Impact of Artificial Intelligence. *Nat. Mach. Intell.*, 2(8):423–425, 2020. 4.2

[51] J. R. Digernes and C. Ditlev-Simonsen. A Workflow-integrated Brain Tumor Segmentation System Based on fastai and MONAI. Master's thesis, University of Bergen, 2022. 4.3

[52] A. A. DiSessa. *Changing Minds: Computers, Learning, and Literacy*. MIT Press, 2000. 4.4

[53] K. Doi. Diagnostic Imaging Over the Last 50 years: Research and Development in Medical Imaging Science and Technology. *Physics in Medicine & Biology*, 51(13):R5, 2006. 3

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. 2.2

[55] B. L. Edlow et al. Data from: 7 Tesla MRI of the ex vivo human brain at 100 micron resolution. `https://doi.org/10.5061/dryad.119f80q`, 2019. 3.1.2, 3.4

[56] European Commission. Annexes to the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 4 2021. 4.4

[57] European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 4 2021. 4.4

[58] R. Feng, M. Badgeley, J. Mocco, and E. K. Oermann. Deep Learning Guided Stroke Management: A Review of Clinical Applications. *Journal of Neurointerventional Surgery*, 10(4):358–362, 2018. 1.1, 3.4

[59] M. Filippi, M. A. Rocca, O. Ciccarelli, N. De Stefano, N. Evangelou, L. Kappos, A. Rovira, J. Sastre-Garriga, M. Tintorè, J. L. Frederiksen, et al. MRI Criteria for the Diagnosis of Multiple Dclerosis: MAGNIMS Consensus Guidelines. *The Lancet Neurology*, 15(3):292–303, 2016. 3.1.2

[60] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. Adversarial Attacks on Medical Machine Learning. *Science*, 363(6433):1287–1289, 2019. 3

[61] K. Fukushima. Neocognitron: A self-organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36(4):193–202, 1980. 2.4, 2.4, 2.1

[62] E. Gibney. This AI researcher is trying to ward off a reproducibility crisis. *Nature*, 577(7788):14, 2020. 5

[63] X. Glorot and Y. Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on*

Artificial Intelligence and Statistics, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 2.2

[64] R. C. Gonzales and P. Wintz. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 2.2.3

[65] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. 1.1, 2.4, 2.4

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, 2020. 2.3, 2.1

[67] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, M. A. Q. C. M. S. B. of Directors, et al. Transparency and Reproducibility in Artificial Intelligence. *Nature*, 586(7829):E14–E16, 2020. 1.1, 4.4

[68] I. S. Haldorsen, N. Lura, J. Blaakær, D. Fischerova, and H. M. Werner. What Is the Role of Imaging at Primary Diagnostic Work-Up in Uterine Cervical Cancer? *Current Oncology Reports*, 21:1–15, 2019. 3.1

[69] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022. 2.2

[70] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 2.2

[71] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2.1

[72] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2.4, 2.2, 4.1

[73] T. Hernandez-Boussard, S. Bozkurt, J. P. Ioannidis, and N. H. Shah. MINIMAR (MINimum Information for Medical AI Reporting): Developing Reporting Standards for Artificial Intelligence in Health Care. *Journal of the American Medical Informatics Association*, 27(12):2011–2015, 2020. 4, 4.4

[74] G. E. Hinton, S. Osindero, and Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006. 2.4

[75] E. Hodneland, S. Kaliyugarasan, K. S. Wagner-Larsen, N. Lura, E. Andersen, H. Bartsch, N. Smit, M. K. Halle, C. Krakstad, A. S. Lundervold, et al. Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer. *Cancers*, 14(10):2372, 2022. 1.2, 1.3, 4.1, 6, 5.4

[76] B. K. Horn. Obtaining Shape From Shading Information. *The Psychology of Computer Vision*, 1975. 2.1

[77] B. K. Horn and B. G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 2.1

[78] S. L. Horowitz and T. Pavlidis. Picture Segmentation by A Tree Traversal algorithm. *Journal of the ACM (JACM)*, 23(2):368–388, 1976. 2.1

[79] J. Howard and S. Gugger. *Deep Learning for Coders with fastai and PyTorch*. O'Reilly Media, 2020. 1.1, 2.4, 2.5

[80] J. Howard and S. Gugger. Fastai: A Layered API for Deep Learning. *Information*, 11(2):108, 2020. 1.1, 5.1

[81] J. Howard and H. Husain. nbdev. https://github.com/fastai/nbdev. Accessed: 2023-02-07. 5.1

[82] J. Howard and S. Ruder. Universal Language Model Fine-tuning for Text Classification. *arXiv preprint arXiv:1801.06146*, 2018. 2.1

[83] H. K. Huang. *PACS and Imaging Informatics: Basic Principles and Applications*. John Wiley & Sons, 2011. 3, 3.3

[84] P. Huang, C. T. Lin, Y. Li, M. C. Tammemagi, M. V. Brock, S. Atkar-Khattra, Y. Xu, P. Hu, J. R. Mayo, H. Schmidt, et al. Prediction of Lung Cancer Risk at Follow-up Screening with Low-dose CT: A Training and Validation Study of a Deep Learning Method. *The Lancet Digital Health*, 1(7):e353–e362, 2019. 3.6

[85] D. H. Hubel and T. N. Wiesel. Receptive Fields of Single Neurones in the Cat's Striate Cortex. *The Journal of Physiology*, 148(3):574, 1959. 2.1

[86] D. H. Hubel and T. N. Wiesel. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. *The Journal of Physiology*, 160(1):106, 1962. 2.1

[87] D. H. Hubel and T. N. Wiesel. Receptive Fields and Functional Architecture of Monkey Striate Cortex. *The Journal of Physiology*, 195(1):215–243, 1968. 2.1

[88] M. Hutson. Artificial Intelligence Faces Reproducibility Crisis. *Science*, 359(6377):725–726, 2018. 4.4

[89] V. Iglovikov and A. Shvets. Ternausnet: U-net with VGG11 Encoder Pre-trained on Imagenet for Image Segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 2.2

[90] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015. 4.1

[91] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: A Self-configuring Method for Deep Learning-based Biomedical Image Segmentation. *Nature methods*, 18(2):203–211, 2021. 1.1, 4.2

[92] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009. 4.1

[93] J. C. Junn, K. A. Soderlund, and C. M. Glastonbury. Imaging of Head and Neck Cancer With CT, MRI, and US. In *Seminars in nuclear medicine*, volume 51, pages 3–12. Elsevier, 2021. 3.1.2

[94] S. Kaliyugarasan, M. H. Dagestad, E. I. Papalini, E. Andersen, J.-A. Zwart, H. Brisby, H. Hebelka, E. Ansgar, K. M. Lagerstrand, and A. S. Lundervold. Multi-Center CNN-based Spine Segmentation From T2w MRI Using Small Amounts of Data. *To appear in the Proceedings of the 20th IEEE International Symposium on Biomedical Imaging (ISBI)*, page 5, 2023. 1.2, 1.3, 4.1, 4.3, 4.4, 5.5, E

[95] S. Kaliyugarasan, M. Kocinski, A. Lundervold, and A. S. Lundervold. 2D and 3D U-Nets for Skull Stripping in a Large and Heterogeneous Set of Head MRI Using fastai. *Proceedings of NIK2020*, 2020. 1.3, 4.1, 5.2

[96] S. Kaliyugarasan and A. S. Lundervold. fastMONAI: A low-code deep learning library for medical image analysis. *Manuscript*, 2023. 1.1, 1.2, 1.3, 4, 4.3, 5.1

[97] S. Kaliyugarasan and A. S. Lundervold. LAB-Net: Lidar and aerial image-based building segmentation using U-Nets. *Nordic Machine Intelligence*, 2(3), 2023. 4, 2.2, 5

[98] S. K. Kaliyugarasan, A. Lundervold, and A. S. Lundervold. Pulmonary Nodule Classification in Lung Cancer From 3D Thoracic CT Scans Using fastai and MONAI. *International Journal of Interactive Multimedia and Artificial Intelligence*, 2021. 1.2, 1.3, 5.3

[99] S. Kapoor and A. Narayanan. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv preprint arXiv:2207.07048*, 2022. 4.4

[100] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour. Deep learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis. *Medical image analysis*, 65:101759, 2020. 4.1

[101] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas. Deep Learning in Robotics: Survey on Model Structures and Training Strategies. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):266–279, 2020. 2.1

[102] S. Karsoliya. Approximating Number of Hidden Layer Neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology*, 3(6):714–717, 2012. 2.4

[103] T. Kattenborn, J. Eichel, and F. E. Fassnacht. Convolutional Neural Networks Enable Efficient, Accurate and Fine-grained Segmentation of Plant Species and Communities from High-resolution UAV Imagery. *Scientific Reports*, 9(1):17656, 2019. 2.2

[104] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Medicine*, 17:1–9, 2019. 1.1, 3.4, 4.3

[105] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment Anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 4.2

[106] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing. Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016. 1.1, 4.4, 5.1

[107] A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features From Tiny Images. Technical report, University of Toronto, Toronto, Ontario, 2009. 2.2

[108] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 60(6):84–90, 2017. 2.1, 2.4, 2.5

[109] J. Kugelman, J. Allman, S. A. Read, S. J. Vincent, J. Tong, M. Kalloniatis, F. K. Chen, M. J. Collins, and D. Alonso-Caneiro. A Comparison of Deep Learning U-Net Architectures for Posterior Segment OCT Retinal Layer Segmentation. *Scientific Reports*, 12(1):14888, 2022. 2.2

[110] M. E. Ladd, P. Bachert, M. Meyerspeer, E. Moser, A. M. Nagel, D. G. Norris, S. Schmitter, O. Speck, S. Straub, and M. Zaiss. Pros and Cons of Ultra-high-field MRI/MRS for Human Application. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 109:1–50, 2018. 3.1.2

[111] P. M. Lauritzen, J. G. Andersen, M. V. Stokke, A. L. Tennstrand, R. Aamodt, T. Heggelund, F. A. Dahl, G. Sandbæk, P. Hurlen, and P. Gulbrandsen. Radiologist-initiated Double Reading of Abdominal CT: Retrospective Analysis of the Clinical Importance of Changes to Radiology Reports. *BMJ quality & safety*, 25(8):595–603, 2016. 1.1

[112] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015. 2.4

[113] 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. *Neural Computation*, 1(4):541–551, 1989. 2.4, 2.4, 2.4, 2.1

[114] P. Lee, S. Bubeck, and J. Petro. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. PMID: 36988602. 1.1

[115] P. Lee, C. Goldberg, and I. Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson, 2023. 1.1

[116] T. C. Lethbridge. Low-code is Often High-code, so We Must Design Low-code Platforms to Enable Proper Software Engineering. In *Leveraging Applications of Formal Methods, Verification and Validation: 10th International Symposium on Leveraging Applications of Formal Methods, ISoLA 2021, Rhodes, Greece, October 17–29, 2021, Proceedings 10*, pages 202–212. Springer, 2021. 1.1

[117] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959. 2.1

[118] D. Leung, X. Han, T. Mikkelsen, and L. B. Nabors. Role of MRI in Primary Brain Tumor Evaluation. *Journal of the National Comprehensive Cancer Network*, 12(11):1561–1568, 2014. 3.1.2

[119] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden. The First Step for Neuroimaging Data Analysis: DICOM to NIfTI Conversion. *Journal of Neuroscience Methods*, 264:47–56, 2016. 3.2

[120] K. Lin, L. Gong, Y. Huang, C. Liu, and J. Pan. Deep Learning-based Segmentation and Quantification of Cucumber Powdery Mildew Using Convolutional Neural Network. *Frontiers in Plant Science*, 10:155, 2019. 2.2

[121] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014. 2.1, 2.1, 2.2, 4.1

[122] C. Liu, C. Gao, X. Xia, D. Lo, J. Grundy, and X. Yang. On the Reproducibility and Replicability of Deep Learning in Software Engineering. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(1):1–46, 2021. 1.1, 4.4

[123] J. Liu, X. Wang, and X.-C. Tai. Deep Convolutional Neural Networks with Spatial Regularization, Volume and Star-Shape Priors for Image Segmentation. *Journal of Mathematical Imaging and Vision*, 64(6):625–645, 2022. 2.5

[124] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 2.2

[125] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2.2, 4.1

[126] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 2.2, 4.1, 4.2

[127] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2.1

[128] J. Lourenco and O. Clark. Clinical Radiology Census Report 2021, 2022. 1.1, 3, 3.3

[129] A. S. Lundervold and A. Lundervold. An Overview of Deep Learning in Medical Imaging Focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. 1.1, 2.1, 2.2.3, 3, 3.4

[130] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. Deep Learning in Remote Sensing Applications: A Meta-analysis and Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 2019. 2.1

[131] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 4.1

[132] G. Marcus. Deep Learning: A Critical Appraisal. *arXiv preprint arXiv:1801.00631*, 2018. 2.1

[133] D. Marr. *Vision: A computational Investigation Into the Human Representation and Processing of Visual Information*. MIT press, 2010. 2.1

[134] D. Marr and T. Poggio. Cooperative Computation of Stereo Disparity: A Cooperative Algorithm is Derived for Extracting Disparity Information from Stereo Image pairs. *Science*, 194(4262):283–287, 1976. 2.1

[135] A. H. Maslow. The psychology of science: A reconnaissance. 1966. 2.5

[136] J. McCarthy et al. What is Artificial Intelligence. 2007. 1.1

[137] C. H. McCollough, J. T. Bushberg, J. G. Fletcher, and L. J. Eckel. Answers to Common Questions About the Use and Safety of CT Scans. In *Mayo Clinic Proceedings*, volume 90, pages 1380–1392. Elsevier, 2015. 3.1.1

[138] W. S. McCulloch and W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, 5:115–133, 1943. 2.4

[139] S. M. McKinney, A. Karthikesalingam, D. Tse, C. J. Kelly, Y. Liu, G. S. Corrado, and S. Shetty. Reply to: Transparency and Reproducibility in Artificial Intelligence. *Nature*, 586(7829):E17–E18, 2020. 4.4

[140] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. Addendum: International evaluation of an AI system for breast cancer screening. *Nature*, 586(7829):E19–E19, 2020. 4.4

[141] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788):89–94, 2020. 4.4

[142] R. Meier, U. Knecht, T. Loosli, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Scientific reports*, 6(1):1–11, 2016. 4.1

[143] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014. 3.3, 4.1, 4.3

[144] D. Merkel. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014. 4.3

[145] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969. 2.4

[146] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. 4.4

[147] MMIV. Workflow-integrated machine learning. `https://mmiv.no/wiml/`, 2021. Accessed: 2023-26-03. 4.3

[148] J. Mongan, L. Moy, and C. E. Kahn Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers, 2020. 4, 4.4, 4.1

[149] E. Montagnon, M. Cerny, A. Cadrin-Chênevert, V. Hamilton, T. Derennes, A. Ilinca, F. Vandenbroucke-Menu, S. Turcotte, S. Kadoury, and A. Tang. Deep Learning Workflow in Radiology: A Primer. *Insights Into Imaging*, 11:1–15, 2020. 1.1, 3, 3.4, 4.1, 4.3

[150] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation Models for Generalist Medical Artificial Intelligence. *Nature*, 616(7956):259–265, 2023. 1.1

[151] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, and M. Maruthappu. Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *BMJ*, 368, 2020. 1.1, 3, 4.4

[152] A. Nair, S. Ramanathan, P. Sathiadoss, A. Jajodia, and D. B. Macdonald. Barriers to Artificial Intelligence Implementation in Radiology Practice: What the Radiologist Needs to Know. *Radiología (English Edition)*, 64(4):324–332, 2022. 4.3

[153] A. Y. Ng. Data-centric AI. `https://datacentricai.org/`, 2022. Accessed: 2023-02-07. 1.1

[154] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh, et al. VinDr-CXR: An Open Dataset of Chest X-rays with Radiologist's Annotations. *Scientific Data*, 9(1):429, 2022. 4.1

[155] E. Niemiec. Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? *Digital Health*, 8:20552076221089079, 2022. 3.4

[156] L. Nilsen. Radiolog-mangel bidrar til uønskede hendelser. https://www.dagensmedisin.no/jobb-og-utdanning-spesialisthelsetjeneste/radiolog-mangel-bidrar-til-uonskede-hendelser/398597, 2017. Accessed: 2023-02-07. 1.1

[157] C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. 4.1

[158] A. Nowogrodzki. The strongest scanners. *Nature*, 563(7729):24–26, 2018. 3.1.2

[159] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall. Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. *arXiv preprint arXiv:1811.03378*, 2018. 2.1

[160] T. O. B. Odden, E. Lockwood, and M. D. Caballero. Physics computational literacy: An exploratory case study using computational essays. *Physical Review Physics Education Research*, 15(2):020152, 2019. 4.4

[161] E. S. of Radiology (ESR) communications@ myESR. org Brady Adrian P. Beets-Tan Regina G. Brkljačić Boris Catalano Carlo Rockall Andrea Fuchsjäger Michael. The Role of Radiologist in the Changing World of Healthcare: a White Paper of the European Society of Radiology (ESR). *Insights into Imaging*, 13(1):100, 2022. 3.3

[162] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2.2

[163] OpenAI. GPT-4 Technical Report, 2023. 1.1, 4.2

[164] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training Language Models to Follow Instructions With Human Feedback, 2022. 1.1

[165] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh. Deep Learning vs. Traditional Computer Vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pages 128–144. Springer, 2020. 2.5

[166] M. Pandey, M. Fernandez, F. Gentile, O. Isayev, A. Tropsha, A. C. Stern, and A. Cherkasov. The Transformational Role of GPU Computing and Deep Learning in Drug Discovery. *Nature Machine Intelligence*, 4(3):211–221, 2022. 2.1

[167] E. Parliament. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC,

Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. *Official Journal of the European Union*, 60(L117):1–175, 2017. 3.4

[168] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2.1

[169] K. Pedersen. 7000 undersøkelser i kø hos røntgenlegene. https://www.bt.no/nyheter/lokalt/i/eLBba/7000-undersokelser-i-ko-hos-rontgenlegene, 2016. 1.1

[170] F. Pérez-García, R. Sparks, and S. Ourselin. TorchIO: A Python Library for Efficient Loading, Preprocessing, Augmentation and Patch-based Sampling of Medical Images in Deep Learning. *Computer Methods and Programs in Biomedicine*, page 106236, 2021. 1.1, 5.1

[171] O. S. Pianykh. *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide*, chapter What Is DICOM?, pages 3–5. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 3.2, 3.3

[172] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire. A Large-scale Study About Quality and Reproducibility of Jupyter Notebooks. In *2019 IEEE/ACM 16th international conference on mining software repositories (MSR)*, pages 507–517. IEEE, 2019. 1.1

[173] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *The Journal of Machine Learning Research*, 22(1):7459–7478, 2021. 4.4, 4.4

[174] C. A. Potter, A. S. Vagal, M. Goyal, D. B. Nunez, T. M. Leslie-Mazwi, and M. H. Lev. CT for Treatment Selection in Acute Ischemic Stroke: A Code Stroke Primer. *Radiographics*, 39(6):1717–1738, 2019. 3.1.1

[175] L. Prechelt. Early Stopping — But When? *Neural networks: Tricks of the Trade: Second Edition*, pages 53–67, 2012. 2.1

[176] N. C. Purandare and V. Rangarajan. Imaging of Lung Cancer: Implications on Staging and Management. *Indian Journal of Radiology and Imaging*, 25(02):109–120, 2015. 3.1.1

[177] P. Puri, N. Comfere, L. A. Drage, H. Shamim, S. A. Bezalel, M. R. Pittelkow, M. D. Davis, M. Wang, A. R. Mangold, M. M. Tollefson, et al. Deep Learning for Dermatologists: Part II. Current Applications. *Journal of the American Academy of Dermatology*, 2020. 1.1

[178] C. Rackauckas, M. Innes, Y. Ma, J. Bettencourt, L. White, and V. Dixit. DiffEqFlux.jl - A Julia Library for Neural Differential Equations. *arXiv preprint arXiv:1902.02376*, 2019. 2.5

[179] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, and A. Ramadhan. Universal Differential Equations for Scientific Machine Learning. *arXiv preprint arXiv:2001.04385*, 2020. 2.5

[180] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. 1.1

[181] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019. 2.2

[182] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. AI in Health and Medicine. *Nature Medicine*, 28(1):31–38, 2022. 1.1, 3.4, 4, 4.4

[183] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 1.1, 4.2

[184] M. Ranzato, Y.-L. Boureau, Y. Cun, et al. Sparse Feature Learning for Deep Belief Networks. *Advances in Neural Information Processing Systems*, 20, 2007. 2.4

[185] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 4.1

[186] P. Riley. Three Pitfalls to Avoid in Machine Learning. *Nature*, 572(7767):27–29, 2019. 2.1

[187] A. Rimmer. Radiologist Shortage Leaves Patient Care at Risk, Warns Royal College. *BMJ: British Medical Journal (Online)*, 359, 2017. 1.1

[188] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nature Machine Intelligence*, 3(3):199–217, 2021. 3

[189] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 4.1

[190] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1.1, 4.2

[191] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2.2, 4.1

[192] F. Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386, 1958. 2.4

[193] S. Ruder. An Overview of Gradient Descent Optimization Algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 2.1

[194] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 3

[195] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536, 1986. 2.4

[196] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2.2

[197] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020. 1.1, 1.1, 2.2.1, 2.4, 2.4

[198] A. Saha, J. J. Twilt, J. S. Bosma, B. van Ginneken, D. Yakar, M. Elschot, J. Veltman, J. Fütterer, M. de Rooij, and H. Huisman. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol). 2022. 4.1, 6

[199] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021. 1.1

[200] M. K. Scheuerman, A. Hanna, and E. Denton. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021. 4.1

[201] F. Schick, C. C. Pieper, P. Kupczyk, H. Almansour, G. Keller, F. Springer, P. Mürtz, C. Endler, A. M. Sprinkart, S. Kaufmann, et al. 1.5 vs 3 Tesla Magnetic Resonance Imaging: A Review of Favorite Clinical Applications for Both Field Strengths-Part 1. *Investigative Radiology*, 56(11):680–691, 2021. 3.1.2

[202] J. Schmidhuber. Deep Learning in Neural Networks: An Overview. *Neural networks*, 61:85–117, 2015. 2.4

[203] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28, 2015. 4.2

[204] B. Settles. Active learning literature survey. 2009. 4.1

[205] C. Shorten and T. M. Khoshgoftaar. A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):1–48, 2019. 2.1

[206] Y. Sim, M. J. Chung, E. Kotter, S. Yune, M. Kim, S. Do, K. Han, H. Kim, S. Yang, D.-J. Lee, et al. Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. *Radiology*, 294(1):199–209, 2020. 3.4

[207] R. G. Smart. The Importance of Negative Results in Psychological Research. *Canadian Psychologist/Psychologie Canadienne*, 5(4):225, 1964. 1.1

[208] S. M. Smith. Fast Robust Automated Brain Extraction. *Human Brain Mapping*, 17(3), 2002. 4.1

[209] R. Smith-Bindman, M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, S. Y. Cheng, E. J. Bowles, J. R. Duncan, R. T. Greenlee, L. H. Kushi, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *Jama*, 322(9):843–856, 2019. 1.1, 3

[210] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, et al. Deep Learning for Electronic Health Records: A Comparative Review of Multiple Deep Neural Architectures. *Journal of Biomedical Informatics*, 101:103337, 2020. 1.1

[211] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 2.1, 4.1

[212] K. Storheim, A. Espeland, L. Grøvle, J. S. Skouen, J. Aßmus, A. Anke, A. Froholdt, L. M. Pedersen, A. J. Haugen, T. Fors, et al. Antibiotic Treatment In Patients With Chronic Low Back Pain and Modic Changes (the AIM study): Study Protocol for a Randomised Controlled Trial. *Trials*, 18(1):1–11, 2017. 4.3

[213] N. H. Strickland. PACS (Picture Archiving and Communication Systems): Filmless Radiology. *Archives of disease in childhood*, 83(1):82–86, 2000. 3, 3.3

[214] P. Suetens. *Fundamentals of Medical Imaging*. Cambridge University Press, 2017. 3, 3.1.1

[215] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852, 2017. 2.2, 4.1

[216] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 2.2

[217] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer Nature, 2022. 2.1, 2.1

[218] M. Takahashi, H. Uematsu, and H. Hatabu. MR imaging at High Magnetic Fields. *European Journal of Radiology*, 46(1):45–52, 2003. 3.1.2

[219] S. Thakar. Radiology Staffing Shortage Seeps into Australian Hospital. https://radiologybusiness.com/topics/medical-practice-management/radiology-staffing-shortage-seeps-australian-hospital, 2018. Accessed: 2023-02-07. 1.1

[220] L. Thesing, V. Antun, and A. C. Hansen. What do AI Algorithms Actually Learn?-On False Structures in Deep Learning. *arXiv preprint arXiv:1906.01478*, 2019. 3

[221] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey. Scientific Machine Learning Benchmarks. *Nature Reviews Physics*, 4(6):413–420, 2022. 4

[222] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. *IEEE Spectrum*, 58(10):50–55, 2021. 3, 4.2

[223] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. LaMDA: Language Models for Dialog Applications, 2022. 1.1, 2.2, 4.2

[224] E. Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette UK, 2019. 1.1

[225] E. J. Topol. High-performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature medicine*, 25(1):44–56, 2019. 1.1, 4

[226] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2.2

[227] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, and A. Madry. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020. 4.1

[228] M. Usman, B.-D. Lee, S.-S. Byon, S.-H. Kim, B.-i. Lee, and Y.-G. Shin. Volumetric Lung Nodule Segmentation Using Adaptive ROI with Multi-view Residual Learning. *Scientific Reports*, 10(1):12839, 2020. 3.6

[229] B. Vachha and S. Y. Huang. MRI with Ultrahigh Field Strength and High-performance Gradients: Challenges and Opportunities for Clinical Neuroimaging at 7 T and Beyond. *European Radiology Experimental*, 5(1):1–18, 2021. 3.1.2, 3.1.2

[230] K. G. van Leeuwen, M. de Rooij, S. Schalekamp, B. van Ginneken, and M. J. Rutten. How Does Artificial Intelligence in Radiology Improve Efficiency and Health Outcomes? *Pediatric Radiology*, pages 1–7, 2021. 1.1, 3.4, 4.3

[231] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij. Artificial Intelligence in Radiology: 100 Commercially Available Products and Their Scientific Evidence. *European Radiology*, 31:3797–3804, 2021. 1.1, 3.4

[232] W. G. Van Panhuis, P. Paul, C. Emerson, J. Grefenstette, R. Wilder, A. J. Herbst, D. Heymann, and D. S. Burke. A systematic review of barriers to data sharing in public health. *BMC public health*, 14(1):1–9, 2014. 1.1, 4.1

[233] G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5(1):48, 2022. 1.1, 4.1, 4.2, 4.4

[234] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 2.1, 2.2

[235] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 4.1

[236] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004. 4.1

[237] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. 4.1

[238] C. Weltens, J. Menten, M. Feron, E. Bellon, P. Demaerel, F. Maes, W. Van den Bogaert, and E. van der Schueren. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiotherapy and Oncology*, 60(1):49–59, 2001. 4.1

[239] B. Whitcher, V. J. Schmid, and A. Thorton. Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software*, 44:1–29, 2011. 3.2

[240] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, et al. Do No Harm: A Roadmap for Responsible Machine Learning for Health Care. *Nature Medicine*, 25(9):1337–1340, 2019. 1.1

[241] J. Willatt, J. A. Ruma, S. F. Azar, N. L. Dasika, and F. Syed. Imaging of Hepatocellular Carcinoma and Image Guided Therapies-how We Do It. *Cancer Imaging*, 17(1):1–10, 2017. 3.1

[242] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing Medical Imaging Data for Machine Learning. *Radiology*, 295(1):4–15, 2020. 4, 4.1

[243] S. Wolfram. What Is a Computational Essay? https://writings.stephenwolfram.com/2017/11/what-is-a-computational-essay/, 2017. Accessed: 2023-31-03. 4.4

[244] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 1.1, 2.2

[245] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith. Bayesian Analysis of Neuroimaging Data in FSL. *Neuroimage*, 45(1):S173–S186, 2009. 4.1

[246] World Health Organization. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. 2022. 1.1, 4.1

[247] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2019. 3.4

[248] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray, et al. Prediction Models for Diagnosis and Prognosis of COVID-19: Systematic Review and Critical Appraisal. *BMJ*, 369, 2020. 3

[249] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 2.2

[250] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. MedMNIST v2-A Large-scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Scientific Data*, 10(1):41, 2023. 4

[251] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357, 2018. 3.6

[252] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 4.1

[253] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 4.1

[254] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 2.4

[255] N. Zha, M. N. Patlas, and R. Duszak Jr. Radiologist Burnout is Not Just Isolated to the United States: Perspectives From Canada. *Journal of the American College of Radiology*, 16(1):121–123, 2019. 1.1, 3

[256] D. Zhang, N. Maslej, E. Brynjolfsson, J. Etchemendy, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, M. Sellitto, E. Sakhaee, Y. Shoham, J. Clark, and R. Perrault. The AI Index 2022 Annual Report, 2022. 1.1, 2.1, 4.4

[257] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4.1

[258] Z. Zhang, Q. Liu, and Y. Wang. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 2.2

[259] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 2.2

[260] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2.2

[261] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 4.1

[262] M. Zinkevich. Rules of Machine Learning: Best Practices for ML Engineering. *URL: https://developers. google. com/machine-learning/guides/rules-of-ml*, 2017. 6, 2

[263] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 2023. 2.2.2

Part II

ARTICLES

PAPER A

FASTMONAI: A LOW-CODE DEEP LEARNING LIBRARY FOR MEDICAL IMAGE ANALYSIS

Kaliyugarasan, Satheshkumar and Lundervold, Alexander Selvikvåg.

Manuscript, April 2023

fastMONAI: a low-code deep learning library for medical image analysis

Satheshkumar Kaliyugarasan   Alexander Selvikvåg Lundervold

Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, HVL
MMIV, Dept. of Radiology, Haukeland University Hospital

April 2023

Open in Colab

Summary

> *"Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do."*(Knuth 1984)

In this work, we present fastMONAI, a low-code Python-based open source deep learning library built on top of fastai (Howard and Gugger 2020b, 2020a), MONAI (Cardoso et al. 2022), and TorchIO (Pérez-García, Sparks, and Ourselin 2021). We created the library to simplify the use of state-of-the-art deep learning techniques in 3D medical image analysis for solving classification, regression, and segmentation tasks. fastMONAI provides users with functionalities to step through data loading, preprocessing, training, and result interpretations.

We've structured the paper as follows: it first discusses the need for the research, then showcases various applications and the library's user-friendliness, followed by a discussion about documentation, usability, and maintainability.

Note that this paper is automatically generated from a Jupyter Notebook available in the fastMONAI GitHub repo: https://github.com/MMIV-ML/fastMONAI. Using the notebook makes it possible to step through the paper's content and reproduce all the text, computations, figures, and results.

Statement of need

Deep learning develops at breakneck speed, with new models, techniques, and tricks constantly appearing. As a result, it is easy to get stuck on something less-than-optimal when using deep learning to solve a particular set of problems while also being in danger of getting lost in minor technical details when constructing models for concrete tasks. Therefore, the fastai deep learning library (Howard and Gugger 2020b, 2020a) provides both a high-level API that automatically incorporates many established best practices and a low-level API in which one can modify details related to model architectures, training strategies, data augmentation, and more.

fastai is a general deep learning library built on top of PyTorch. However, medical imaging has a variety of domain-specific demands, including medical imaging formats, data storage and transfer, data labeling procedures, domain-specific data augmentation, and evaluation methods. MONAI Core (Cardoso et al. 2022) and TorchIO (Pérez-García, Sparks, and Ourselin 2021) target deep learning in healthcare imaging, incorporating multiple best practices. MONAI Core, the primary library of Project MONAI, is built on top of PyTorch and provides domain-specific functionalities for medical imaging, including network architectures, metrics, and loss functions. TorchIO is a Python-based open-source library for efficiently loading, preprocessing, and augmenting 3D medical images.

Three key features impacting the performance of a deep learning system are the network architecture, training methods, and data (Woo et al. 2023). Our combination of fastai, MONAI Core, and TorchIO into fastMONAI, together with custom-made modules like our `MedDataset`, makes it possible to easily construct, train, and use powerful models with a range of different architectures for a variety of medical imaging tasks, while using established best practices for training, for reading data, for performing data augmentation, and for other domain-specific capabilities incorporated into these three libraries.

The library is developed at The Mohn Medical Imaging and Visualization Centre (MMIV), which is part of the Department of Radiology at Haukeland University Hospital. One of the center's key objectives is to develop new quantitative methods for high-field MRI, CT, and hybrid PET/CT/MR in preclinical and clinical settings, aiming to improve decision-making and patient care. fastMONAI supports such efforts by easing the entry for new practitioners into medical AI and making it possible to quickly construct good baseline models while still being flexible enough to enable further optimizations.

Using fastMONAI

In this section, we will explore how to use our library. In fastMONAI's online documentation https://fastmonai.no, multiple tutorials cover classification, regression, and segmentation tasks.

Classification

After installing the library, the first step is to import the necessary functions and classes. For example, the following line imports all of the functions and classes from the fastMONAI library:

```
from fastMONAI.vision_all import *
```

Downloading external data

To demonstrate the use of fastMONAI. we download the NoduleMNIST3D dataset from MedMNIST v2 (Yang et al. 2023), a dataset containing lung nodules with labels indicating whether the nodules are benign (b) or malignant (m):

```
df, _ = download_NoduleMNIST3D(max_workers = 8)
```

Inspecting the data

Let's look at how the processed DataFrame is formatted:

```
print(df.head(1).to_markdown())
```

img_path	labels	is_val
../data/NoduleMNIST3D/train_images/0_nodule.nii.gz	b	False

In fastMONAI, various data augmentation techniques are available for training vision models, and they can also optionally be applied during inference. The following code cell specifies a list of transformations to be applied to the items in the training set. The complete list of available transformations in the library can be found at https://fastmonai.no/vision_augment.

```
item_tfms = [PadOrCrop(size = 28), RandomAffine(degrees = 35, isotropic = True),
             ZNormalization()]
```

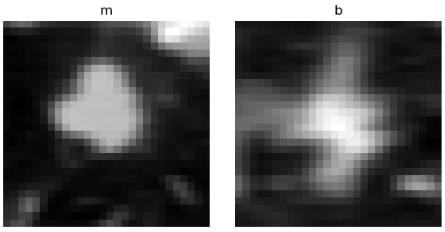
Before feeding the data into a model, we must create a `DataLoaders` object for our dataset. There are several ways to get the data in `DataLoaders`. In the following line, we call the `ImageDataLoaders.from_df` factory method, which is the most basic way of building a `DataLoaders`.

Here, we pass the processed DataFrame, define the columns for the images `fn_col` and the labels `label_col`, some transforms `item_tfms`, voxel spacing `resample`, and the batch size `bs`.

```
dls = MedImageDataLoaders.from_df(df, fn_col = 'img_path', label_col = 'labels',
                                  item_tfms = item_tfms, resample = 1, bs = 64)
```

We can now take a look at a batch of images in the training set using `show_batch` :

```
dls.show_batch(max_n = 2, anatomical_plane = 2)
```



Choosing a loss function

Class imbalance is a common challenge in medical datasets, and it is something we're facing in our example dataset:

```
print(df.labels.value_counts())
```

```
b    986
m    337
```

There are multiple ways to deal with class imbalance. A straightforward technique is to use balancing weights in the model's loss function, i.e., penalizing misclassifications for instances belonging to the minority class more heavily than those of the majority class.

```
train_labels = df.loc[~df.is_val]['labels'].tolist()
class_weights = get_class_weights(train_labels)
print(class_weights)
```

```
tensor([0.6709, 1.9627])
```

```
loss_func = CrossEntropyLossFlat(weight = class_weights)
```

We're now ready to construct a deep learning classification model.

Create and train a 3D deep learning model

We import a classification network from MONAI and configure it based on our task, including defining the input image size, the number of classes to predict, channels, etc.

```
from monai.networks.nets import Classifier

model = Classifier(in_shape = [1, 28, 28, 28], classes = 2,
                   channels = (8, 16, 32, 64), strides=(2, 2, 2))
```

Then we create a `Learner`, which is a fastai object that combines the data and our defined model for training.

```
learn = Learner(dls, model, loss_func = loss_func, metrics = accuracy)
```

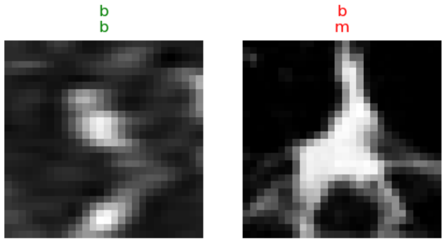
```
learn.fit_one_cycle(4)
```

epoch	train_loss	valid_loss	accuracy	time
0	0.568710	0.464355	0.780303	00:02
1	0.532495	0.512985	0.818182	00:02
2	0.480088	0.471735	0.829545	00:02
3	0.438136	0.436847	0.825758	00:02

Note: Small random variations are involved in training CNN models. Hence, when running the notebook, you may see different results.

With the model trained, let's look at some predictions on the validation data. The `show_results` method plots instances, their target values, and their corresponding predictions from the model.

```
learn.show_results(max_n = 2, anatomical_plane = 2)
```

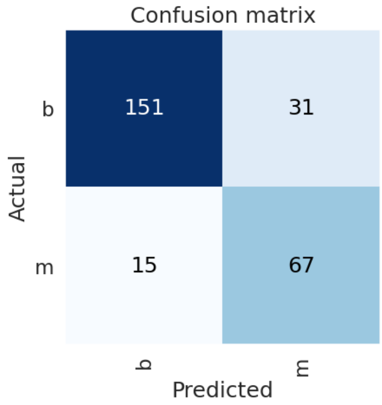


Model evaluation and interpretation

Let's look at how often and for what instances our trained model becomes confused while making predictions on the validation data:

```
interp = ClassificationInterpretation.from_learner(learn)
```

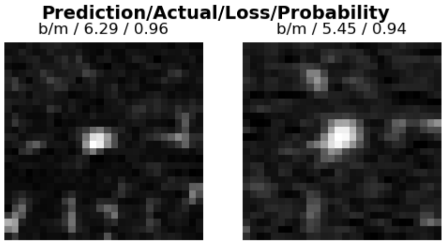
```
interp.plot_confusion_matrix()
```



	b	m
b	151	31
m	15	67

Here are the two instances our model was most confused about (in other words, most confident but wrong):

```
interp.plot_top_losses(k = 2, anatomical_plane = 2)
```



Improving results using test-time augmentation

Test-time augmentation (TTA) is a technique where you apply data augmentation transforms when making predictions to produce average output. In addition to often yielding better performance, the variation in the output of the TTA runs can provide some measure of its robustness and sensitivity to augmentations.

```
preds, targs = learn.tta(n = 4);
print(accuracy(preds, targs))
```

```
TensorBase(0.8371)
```

Semantic segmentation

In the following, we look at another computer vision task while also taking a closer look at the fastMONAI library. Our task will be *semantic segmentation*, and we'll use the IXI Tiny dataset provided by TorchIO (a small version of the IXI dataset (Alansary et al., n.d.)) with 566 3D brain MRI scans. In semantic segmentation, a class label is assigned to each pixel or voxel in an image, in this case, distinguishing brain tissue from non-brain tissue, i.e., skull-stripping or brain extraction.

```
STUDY_DIR = download_ixi_tiny(path = '../data')
```

```
df = pd.read_csv(STUDY_DIR/'dataset.csv')
```

Inspecting the data

The fastMONAI class `MedDataset` can automatically extract and present valuable information about your dataset:

```
med_dataset = MedDataset(path = STUDY_DIR/'image', reorder = True, max_workers = 6)
```

```
data_info_df = med_dataset.summary()
print(data_info_df.head().to_markdown())
```

	dim0	dim1	dim2	vx0	vx1	vx2	orient	example_path	total
0	44	55	83	4.13	3.95	2.18	RAS+	../data/..IXI002..	566

```
resample, reorder = med_dataset.suggestion()
print(resample, reorder)
```

```
[4.13, 3.95, 2.18] True
```

We can get the largest image size in the dataset with the recommended resampling:

```
img_size = med_dataset.get_largest_img_size(resample)
print(img_size)
```

```
[44.0, 55.0, 83.0]
```

In this case, we choose the following size as some network architectues requires the tensor to be divisible by 16.

```
size = [48, 48, 96]
```

```
item_tfms = [PadOrCrop(size),
             RandomAffine(scales = 0.1, degrees = 5, p = 0.5), RandomFlip(p = 0.5),
             ZNormalization()]
```

Loading the data

As we mentioned earlier, there are several ways to get the data in `DataLoaders`. In this section, let's build the data loaders using `DataBlock`. Here we need to define what our input and target should be (`MedImage` and `MedMaskBlock` for segmentation), how to get the images and the labels, how to split the data, item transforms that should be applied during training, reorder voxel orientations, and voxel spacing. Take a look at fastai's documentation for DataBlock for further information: https://docs.fast.ai/data.block.html#DataBlock.

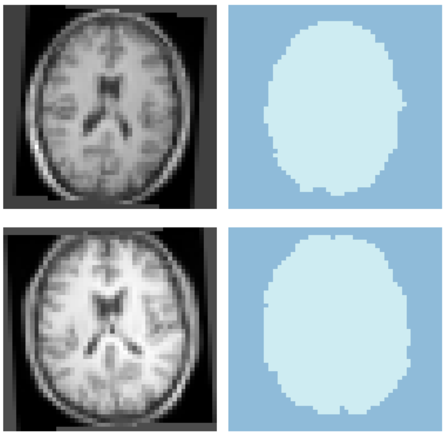
NB: It is crucial to select an appropriate splitting strategy. For example, one should typically avoid having data from the same patient in both the training and the validation or test set. However, in the IXI data set this is not an issue, as there is only one image per patient.

```
dblock = MedDataBlock(blocks=(ImageBlock(cls = MedImage), MedMaskBlock),
                      splitter=RandomSplitter(valid_pct = 0.2, seed = 42),
                      get_x = ColReader('t1_path'), get_y = ColReader('labels'),
                      item_tfms = item_tfms, reorder = reorder, resample = resample)
```

Now we pass our processed DataFrame and the batch size (bs) to create a `DataLoaders` object:

```
dls = dblock.dataloaders(df, bs = 8)
```

```
dls.show_batch(max_n = 2, anatomical_plane = 2)
```



Network architectures and loss functions

You can import various models and loss functions directly from MONAI Core, as shown below:

```
from monai.networks.nets import UNet, AttentionUnet
from monai.losses import DiceLoss, DiceFocalLoss

loss_func = CustomLoss(loss_func=DiceFocalLoss(sigmoid = True))

model = AttentionUnet(spatial_dims = 3, in_channels = 1, out_channels = 1,
                channels = (16, 32, 64, 128), strides = (2, 2, 2))
```

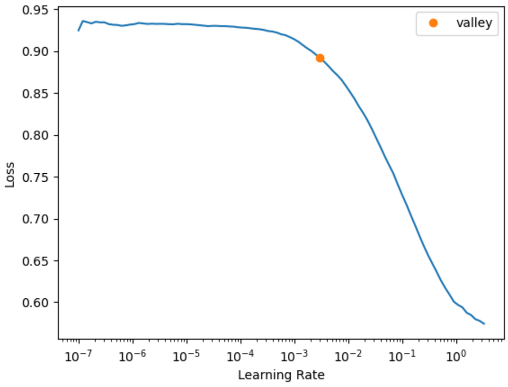
In this task, we use the Ranger optimizer (Wright 2019), a optimization algorithm that combines RAdam and Lookahead.

```
learn = Learner(dls, model, loss_func = loss_func, opt_func = ranger,
                metrics = [binary_dice_score, binary_hausdorff_distance])
```

Finding a good learning rate

We used the default learning rate before, but we might want to find a better value. For this, we can use the learning rate finder of fastai:

```
lr = learn.lr_find()
```



Training the model

Now we can train the model. This time we use `fit_flat_cos` method, which is a better learning rate policy for the Ranger optimzer:

```
learn.fit_flat_cos(2, lr.valley)
learn.save('model-1')
```

epoch	train_loss	valid_loss	dice_score	hausdorff_distance	time
0	0.448831	0.355411	0.932139	7.755924	00:16
1	0.337796	0.278777	0.959338	5.635414	00:15

Exporting and sharing models

We can export the model and share both the trained weights and the learner on HuggingFace and use tagging for marked version release. Version control for shared models is essential for tracking changes and being able to roll back to previous versions if there are any issues with the latest model in production.

```
learn.export('models/export.pkl')
store_variables('models/vars.pkl', size, reorder, resample)
```

Documentation, usability, and maintainability

We have written the entire fastMONAI library using nbdev, a tool for exploratory programming that allows you to write, test, and document a Python library in Jupyter Notebooks. fastMONAI contains several practical tools to ensure the software's user-friendliness.

fastMONAI comes with a documentation page https://fastmonai.no and step-by-step tutorials on how to use the software for various medical imaging tasks (e.g., classification, regression, and segmentation). Tests are written directly in notebooks, and continuous integration with GitHub Actions runs the tests on each push, making software development easier with multiple collaborators.

To ease further extensions of our library through contributions, we have added a short guide on how to contribute to the project. As mentioned, this paper is written as a notebook and automatically converted to a markdown file. The latest version is always available on GitHub.

Research projects using fastMONAI

The fastMONAI library has been used for various medical imaging tasks, including predicting brain age using T1-weighted scans in (Kaliyugarasan, Lundervold, and Lundervold 2020), skull-stripping in (Kaliyugarasan et al. 2020), pulmonary nodule classification from CT images in (Kaliyugarasan, Lundervold, and Lundervold 2021), and tumor segmentation in cervical cancer from multi-parametric pelvic MRI in (Hodneland, Kaliyugarasan, et al. 2022). Recently, it was also used for vertebra segmentation in a multi-center study (Kaliyugarasan, Lundervold, et al. 2023).

Acknowledgments

The Trond Mohn Research Foundation supported our work through the project *"Computational medical imaging and machine learning - methods, infrastructure and applications"* at the Mohn Medical Imaging and Visualization Centre, grant number BFS2018TMT07, and a grant from the Western Norway Regional Health Authority (Helse Vest RHF), project F-12532.

References

Alansary, Amir et al. n.d. "IXI Dataset." http://brain-development.org/ixi-dataset/; Biomedical Image Analysis Group, Imperial College London.

Cardoso, M Jorge, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, et al. 2022. "MONAI: An Open-Source Framework for Deep Learning in Healthcare." *arXiv Preprint arXiv:2211.02701*.

Hodneland, E, S Kaliyugarasan, et al. 2022. "Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer." *Cancers* 14 (10): 2372. https://doi.org/10.3390/cancers14102372.

Howard, Jeremy, and Sylvain Gugger. 2020a. *Deep Learning for Coders with fastai and PyTorch*. O'Reilly Media.

———. 2020b. "Fastai: a layered API for deep learning." *Information* 11 (2): 108. https://doi.org/10.3390/info11020108.

Kaliyugarasan, S, M Kocinski, A Lundervold, and AS Lundervold. 2020. "2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using fastai." *Proceedings of the NIK2020*.

Kaliyugarasan, S, A Lundervold, and AS Lundervold. 2020. "Brain age versus chronological age: A large scale MRI and deep learning investigation." *Proceedings of European Congress of Radiology-ECR 2020*. https://doi.org/10.26044/ecr2020/C-05555.

———. 2021. "Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using fastai and MONAI." https://doi.org/10.9781/ijimai.2021.05.002.

Kaliyugarasan, S, AS Lundervold, et al. 2023. "Multi-Center CNN-based spine segmentation from T2w MRI using small amounts of data." *To Appear in the Proceedings of the 20th IEEE International Symposium on Biomedical Imaging (ISBI)*, 5.

Knuth, D. E. 1984. "Literate Programming." *The Computer Journal* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Pérez-García, Fernando, Rachel Sparks, and Sébastien Ourselin. 2021. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning." *Computer Methods and Programs in Biomedicine*, 106236. https://doi.org/10.1016/j.cmpb.2021.106236.

Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders." *arXiv Preprint arXiv:2301.00808*. https://doi.org/10.48550/arXiv.2301.00808.

Wright, Less. 2019. "Ranger - a Synergistic Optimizer." *GitHub Repository*. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer; GitHub.

Yang, Jiancheng, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. "MedMNIST V2-a Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification." *Scientific Data* 10 (1): 41. https://doi.org/10.1038/s41597-022-01721-8.

PAPER B

2D AND 3D U-NETS FOR SKULL STRIPPING IN A LARGE AND HETEROGENEOUS SET OF HEAD MRI USING FASTAI

Kaliyugarasan, Satheshkumar, Kociński, Marek, Lundervold, Arvid and Lundervold, Alexander Selvikvåg.

In Proceedings of the of the 33rd Norwegian Informatics Conference (NIK), 23 November 2020

2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using `fastai`

Satheshkumar Kaliyugarasan[1,2,*], Marek Kociński[1,3,4,*], Arvid Lundervold[1,3,*], Alexander Selvikvåg Lundervold[1,2,*], for the Alzheimer's Disease Neuroimaging Initiative[**], and for the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[***]

[1]Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland University Hospital, Bergen, Norway
[2]Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[3]Dept. of Biomedicine, University of Bergen, Norway
[4]Institute of Electronics, Lodz University of Technology, Poland
[*]All authors contributed equally to the work

[**]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
`http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

[***]Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database. The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at `www.aibl.csiro.au`.

Abstract

Skull stripping in brain imaging is the removal of the parts of images corresponding to non-brain tissue. Fast and accurate skull stripping is a crucial step for numerous medical brain imaging applications, e.g. registration, segmentation and feature extraction, as it eases subsequent image processing steps. In this work, we propose and compare two novel skull stripping methods based on 2D and 3D convolutional neural networks trained on a large, heterogeneous collection of 2777 clinical 3D T1-weighted MRI images from 1681 healthy subjects. We investigated the performance of the models by testing them on 927 images from 324 subjects set aside from our collection of data, in addition to images from an independent, large brain imaging study: the IXI dataset ($n = 556$). Our models achieved mean Dice scores higher than 0.978 and Jaccard indices higher than 0.957 on all tests sets, making predictions on new unseen brain MR images in approximately 1.4s for the 3D model and 12.4s for the 2D model. A preliminary exploration of the models' robustness to variation in the input data showed favourable results when compared to a traditional, well-established skull stripping method. With further research aimed at increasing the models' robustness, such accurate and fast skull stripping methods can potentially form a useful component of brain MRI analysis pipelines.

1 Introduction

Magnetic resonance imaging of the brain

Magnetic resonance imaging (MRI) is a medical imaging technology (modality) used in radiology to acquire information in space and time about structure (anatomy) and

This paper was presented at the NIK-2020 conference; see `http://www.nik.no`*.*

function (physiology) of tissues and organs in the body. MRI scanners use a combination of strong magnetic fields, magnetic field gradients for spatial encoding and decoding of nuclear spin populations, typically protons (e.g. water) in different chemical and microstructural environments, radio waves, and image reconstruction algorithms working in complex-valued Fourier space. This is used to generate 2D, 3D, 3D+time, or even higher dimensional images of organs, providing information about tissue states and physiological and biochemical processes. Among the most frequent organs subject to MRI examinations is the brain. There are several reasons for this: (i) MRI measurements can collect unsurpassed rich and detailed soft tissue information from the living brain in health and disease with little risk for the patient, and at multiple times during a disease process; (ii) compared to most other parts of the body the brain is an organ for which invasive biopsies (tissue samples) are rarely indicated, for obvious reasons; (iii) the brain within the skull can be kept rather stationary in the head coil during MR measurement time (total examination time is usually 15 - 45 min) in contrast to e.g. the beating heart or abdominal organs that move due to respiration and pulsations causing displacements and movement artifacts that are challenging to correct for, and finally (iv) most of the new MRI measurement techniques (e.g. high resolution structural MRI, diffusion MRI and functional MRI) and advanced image analysis developments tend to first enter the brain and neuro-imaging field before being adapted and applied to other organs.

Deep learning in brain imaging

Recent years' surge of interest in image analysis approaches based on *deep learning* is a case in point [1]. Considerable advances in computers' ability to extract meaningful, actionable information from complicated and heterogeneous datasets have resulted in remarkable achievements in general computer vision, natural language processing, data synthesis, sequence analysis, robotics, the analysis of tabular structured datasets, and more. Driven by these advances, the field of artificial intelligence is experiencing a tremendous amount of attention from researchers, industry, funding agencies, government and entrepreneurs, leading to rapid progress in methods, applications and products. Artificial intelligence in medicine has a long history, dating back to at least the early 1970s[1], but the field hasn't yet had a broad impact on medical practice [3]. Recently, the possibilities of using deep learning on medical data has proven to be highly potent, leading to a torrent of publications across many medical disciplines: radiology, psychiatry, dermatology, pathology, ophthalmology, cardiology, electronic health records, drug discovery, genome sequencing, and much more. See the continuously updated review `https://greenelab.github.io/deep-review`.[2]

What is skull stripping?

Skull stripping, also called brain extraction, is the task of extracting the cerebrum and the cerebellum, including cerebrospinal fluid (CSF) in the subarachnoid space from a given 3D MRI head acquisition (cf. Fig. 1). The brainstem is cut according to a specified level (e.g. distal part of medulla oblongata), assuming this level of the central nervous system is located within the field of view of the image. See the white arrow in Fig. 2 d) for an illustration. Inclusion of extra-dural tissue, e.g. skull, scalp, muscle or fat, or exclusion of brain parenchyma proper, e.g. cuts into gray matter or white matter, are considered skull stripping failures.

[1] e.g. the Mycin system of [2] aimed at identifying bacterial infections and recommending antibiotics

[2] A soon-to-be-updated published version of the survey from 2018 is available in [4]

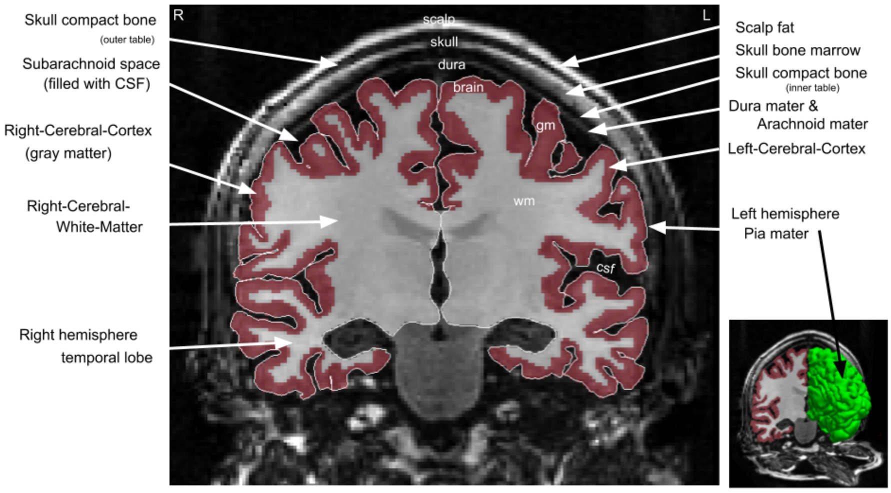


Figure 1: *Anatomy of the head related to the brain extraction task.* A coronal slice from a 3D T1-weighted (T1w) MRI recording from the head showing the different anatomical structures relevant to the segmentation task of skull stripping or brain extraction (data from [5]). Fully automated segmentation of brain (ribbon) including gray matter (gm) and white matter (wm) of the left and right hemisphere and the outer pial boundary of the brain (white continuous tracing and the surface rendering in the small insert) was performed using `Freesurfer v.7.1.1`. CFS = cerebrospinal fluid. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure1`.

Skull stripping is important

Skull stripping is essentially *a region of interest (ROI) segmentation procedure* for subsequent analysis of structural and functional image-derived properties, spatially restricted to the brain, the brainstem (midbrain, pons, medulla oblongata) and the cerebellum. Considering signal intensities, several tissues outside the skull will have intensity distributions that overlap with principal tissue types within the brain. E.g. skeletal muscle in the head have very similar signal intensities in T1w MRI acquisitions to those observed in cerebral gray matter, and blood perfusion time courses or water diffusion properties outside the skull might have similar shape or characteristics as observed within the brain. Thus, for visualization purposes and for quantification (e.g. mean value of an imaging-derived parameter with in the brain) a skull stripping procedure is essential. Moreover, a spatially meaningful restriction of a 2D, 3D or 4D (multispectral 3D or 3D+time) image will help subsequent segmentation algorithms in further spatial refinement and increased anatomical and functional granularity within the brain (e.g. tissue classification in health and disease, or functional connectivity analysis from fMRI recordings assuming all nodes in a network graph are located within the brain, or a sub-region of the brain).

Skull stripping is difficult

There are many sources of difficulty for brain MRI image analysis methods, ranging from scanner and acquisition protocol variation, to subject motion and varying head position in the coil. One important challenge is the presence of a *bias field*. This is is usually perceived as a low-frequency, smooth variation of intensities across a slice image that degrades the MRI recording. The same tissue occurring at different locations within the image can have different signal intensity, invalidating the piecewise constant property of ideal images. Such MRI bias field is caused by an improper image acquisition process, such as radio frequency coil (B_1) non-uniformity or inhomogeneity of the main magnetic field (B_0), this being more prevalent in older MRI scanners or in ultra high-field ($B_0 \geq 7$ T) scanners. Trained radiologist are hardly influenced by this, as they easily compensate for this non-biological intensity variation in image regions. However, the bias field can pose a difficulty for quantitative image analysis algorithms assuming a spatially invariant relation between signal intensity (gray level) distribution and underlying tissue type or state. In the context of skull stripping and brain segmentation, a bias field correcting algorithm is therefore typically applied as a preprocessing step. See Fig. 2 for an example.

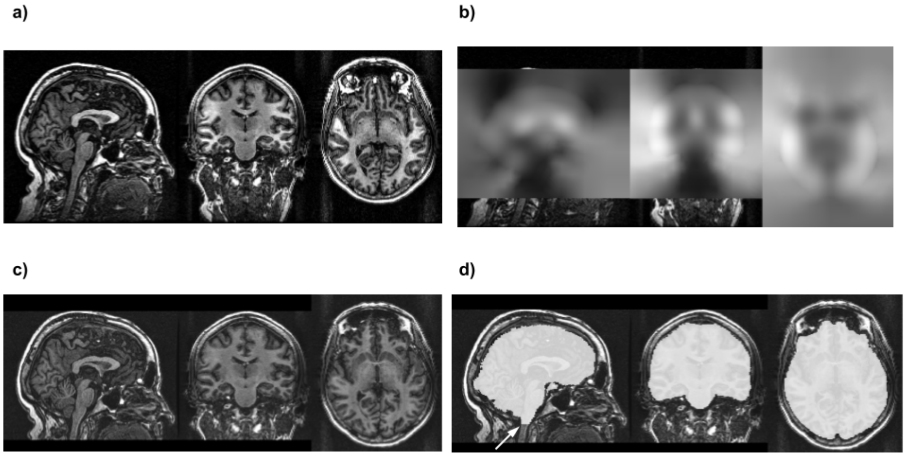


Figure 2: Bias field correction and skull stripping using `fsl_anat` (same subject as in Fig. 1). a) Original acquisition with substantial signal intensity inhomogeneity due to the presence of a bias field. b) The bias field estimate (low-frequency and smooth intensity variation across the image). c) Bias field corrected (i.e. suppressed) image. d) Skull stripping of the bias field corrected image. Arrow indicate the cut of the brainstem, defining the lower boundary of the brain.

Related work

As it is such a fundamental task in brain image analysis there has been a lot of research into skull stripping since the advent of brain MRI image analysis, leading to many proposed methods. These can be roughly categorized into machine learning- and conventional non-machine learning-based approaches. Among conventional methods there's a wide variety of approaches, based on surfaces, morphology, image intensity, templates, or hybrids of these, resulting in a number of well-established, frequently used skull stripping tools in brain image analysis pipelines [6], e.g. the Brain Extraction Tool (BET) [7] of the FMRIB Software Library (FSL), v.6.0 [8], `antsBrainExtraction` from Advanced Normalization Tools (ANTs) [9] and the `3dSkullStrip` tool in AFNI [10]. The machine learning approaches are either based on "classical" machine learning models, e.g. SVMs, region growing, active contours, or based on deep neural networks. It is the latter category that our own work belongs. Two recent illustrative examples of related approaches are presented below.

In [11] the authors developed an automated skull stripping algorithm called **HD-BET** that works for pre-contrast T1w, post-contrast T1w, T2w and FLAIR sequences. Their three-dimensional U-Net-like CNN was trained on 6.586 MR images from 1568 exams of 372 patients collected at 25 different institutions in the EORTC-2610 study. As ground truth brain masks they used BET as a starting point then had a radiologist do visual inspection and corrections (i.e. a single rater). During training, the images were resampled to isotropic spacing of 1.5mm^3 and patches of size 128^3 voxels were randomly sampled from the four different input modalities before being fed to the model. They used a relatively large set of data augmentation techniques: randomly mirroring the image patches along all axes, scaling, rotation and elastic deformations, gamma augmentation, adding additive Gaussian noise, and Gaussian blurring. They scored their model on five independent test sets: one created using the data from 12 institutions in the EORTC-2610 study not present in their training data, and the three openly available datasets LPBA40 from LONI, NFBS and CC-359, for which manually constructed ground truth masks are available. On T1w images from the EORTC-2610 study, their model had a median Dice score of 97.6 (97.0-98.0 IQR) and a median Hausdorff distance of 3.3 (2.2-3.3 IQR). On the three openly available datasets their model obtained a Dice score of 97.5 (17.4-97.7), 98.2 (98.0-98.4), 96.9 (96.7-97.1), respectively, when compared to the provided ground truth reference masks.

The **CompNets** of [12] are multi-pathway two-dimensional U-Net-like models with an embedded W-Net-like component [13], tasked with extracting information from both the brain and non-brain tissue in the input images. Their models were trained on T1w images from 406 subjects aged 18-96 from the OASIS dataset, using the brain masks provided with the OASIS dataset release as ground truth labels. All images were of size 256^3, and their models were trained using 2D slices of the 3D images, with no data augmentation. After making predictions, the masks for each slice were stacked into 3D images. No postprocessing of the resulting predicted brain masked was performed. In a two-fold cross-validation setup were the OASIS subjects are equally divided into two chunks for training and testing, their best model achieved an average Dice score of 98.27 ± 0.30.

Main contributions of our work

1. We construct high-performing skull-strip models from a large, heterogeneous dataset sourced from seven different brain imaging studies. Our results compare favourably with other state-of-the-art models based on deep learning, although direct comparisons between methods are difficult because of the lack of an agreed upon ground truth. This is an issue we discuss in our work.

2. With a novel combination of the `MONAI` deep learning library and our own extension of the `fastai` library to 3D problems, we are able to use multiple interesting state-of-the-art techniques for the construction and training of models.

3. We evaluate the performance of our models on data completely unseen during model construction. Some of which were gathered by a brain imaging study using different combinations of scanners and scanning protocols than those represented in the training data.

4. Once a skull stripping approach reaches a certain average performance level, then arguably the robustness to variation in the input data becomes more important than increased average performance. Our work indicates that CNN-based approaches to skull stripping have some robustness advantages over traditional methods.

5. The data used in our study is available to researchers through various project websites (linked below), easing reproductions and comparisons with other skull stripping methods.

2 Methods and materials

Image datasets

We compiled a large collection of T1w images of healthy volunteers from a number of different data sources[3]: ADNI, AIBL, IXI, PPMI, SLIM, Calgary-Campinas and SALD. This is a highly heterogeneous collection, involving a large number of subjects, scanners and scanner protocols, image sizes and voxel spacings, making it a challenge for any model to make predictions, but also leading to models that are more robustness to such variation. The studies were approved by the relevant Institutional Review Boards at each site and informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. Part of the data material used was sourced from the ADNI database. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of Mild Cognitive Impairment (MCI) and early Alzheimer's Disease (AD) [14]. We also used data collected by the AIBL study group. AIBL study methodology has been reported previously [15].

Note that we selected the subjects that were marked as healthy throughout all these longitudinal studies.

Image preprocessing and label generation

The steps used to automatically construct ground truth labels were performed using a set of well-established, validated tools. There are no manual steps in this process, making it easy to scale our approach to a large number of images. The DICOM recordings were first converted to NIfTI data format using `dicom2niix` (v1.0.20190902, [16]). To reduce the effect of scanner variation, we performed bias field correction, before producing masks indicating the location of the brain. These last two steps were done using a combination of multiple tools from the FMRIB Software Library (FSL) v.6.0 [8], collected in the `fsl_anat` pipeline[4]: (i) reorientation to match the MNI152 standard template orientation using `reorient2std` in FSL, (ii) bias field correction using `FAST` [17], (iii) linear and nonlinear registration to standard MNI152 space using `FLIRT` and `FNIRT` [18, 19], from which the brain was extracted [7]. The entire set of preprocessing steps takes on average less than 10 minutes per volume on a standard workstation computer (e.g. on an Intel Core i7-7700K CPU running Ubuntu 18.04 GNU/Linux). Finally, all volumes were resampled to isotropic $1.0 \times 1.0 \times 1.0$ mm^3 voxel size with the use of the `Convert3D Tool`.

The preprocessed images and ground truth masks were used to create training and testing datasets. Our 2D and 3D setups were based on exactly the same underlying subjects and images, placed in common training and test sets. The training dataset for our 3D model contained 2791 NIfTI files, while the two test sets, tes and IXI, consisted of 934 and 561 images, respectively. For the 2D approach, each 3D volume ($\sim$ 170 axial slices) was split into a set of 2D axial cross-sections. The total number of image files used to train the 2D model was then 469.116, while the two test datasets contained 157.036 and 95.520 image files, respectively.

[3] Links to all the data sources used in this work can be found here: `https://github.com/MMIV-ML/Skull-stripping-NIK2020`

[4] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat`

	2D U-Net	**3D U-Net**
Optimizer	Adam	Adam
Base learning rate	0.0001	0.01
Loss function	Binary Cross-Entropy	Based on the Jaccard Index
Image size	128 x 128	160 x 160 x 92
Data augmentation	Random rotation [$-15°$, $15°$] and scaling [1.0, 1.05]	Random rotation [$-10°$, $10°$] and scaling [1.0, 1.1]
Dropout	0.5	0.5
Weight decay	0.01	0.01
Batch size	128	8
GPU	NVIDIA Titan RTX 24GB	4 x NVIDIA Tesla V 1000 32GB

Table 1: Experimental settings for our 2D and 3D U-Net models.

Constructing and training the 2D and 3D models

We used two different U-Net models in this work: (i) A dynamic 2D U-Net using a ResNet-34 model pre-trained on the ImageNet dataset for image feature extraction (encoder) and PixelShuffle [20] with ICNR initalization [21] for upsampling (decoder), implemented in the PyTorch-based `fastai v1`; (ii) a 3D U-Net implemented using MONAI, a PyTorch based library for deep learning in healthcare imaging, and trained using our own extension of the `fastai` library. The computer vision implementations of the `fastai` library are mostly tailored to 2D imaging. We adapted the library to 3D MR images by constructing new data loaders and data augmentation capabilities, as well as adapting various 2D-specific functionality in the `fastai` library. This enables the use of custom 3D CNNs while still supporting the highly impactful training techniques of `fastai`. This includes the learning rate finder to find the optimum learning rate and the one-cycle policy (e.g., learning rate changes during the training, related to what is called superconvergence [22]). See Table 1 for details about experimental settings.

Performance evaluation

We evaluated the models using the two different test sets described above: (i) data put aside from the training data repositories, 10% of each, making sure there were no subjects appearing in both training and test and controlling for age by stratification over age groups; (ii) the IXI dataset, i.e. data from a completely independent study of 561 subjects, simulating a more realistic use-case for the models. For the hold-out set in (i), we report both the overall results and the results on each repository.

As performance metrics we used the Sørensen-Dice similarity coefficient (DSC) and the Jaccard index (Jacc), measuring the degree of overlap between the ground truth masks generated by FSL and the model predictions. We also used the Hausdorff distance (Haus) between the two masks as a metric. The DSC is the mean overlap of the masks, while

Jacc is the union overlap, and Haus is a measure of extreme deviation between the masks:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}, \qquad \text{Jacc} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}, \qquad \text{Haus} = \max\{(h(X,Y), h(Y,X)\}$$

where $h(X,Y)$ identifies the voxel $x \in X$ that is farthest from any voxel of Y and measures the distance from x to its nearest neighbor in Y. This means that $h(X,Y)$ first looks for the nearest voxel in Y for every voxel in X, and then the largest of these values are taken as the distance, which is the most mismatched point of X. Similarly for $h(Y,X)$, meaning that Haus(X,Y) is able to measure the degree of mismatch between ground truth X and prediction Y from the distance of the point of X that is farthest from any point of Y, and vice versa.

A data filter

While looking at some training images and their corresponding ground truth labels, we observed a few images that were incorrectly labeled as shown in Fig. 3. In order to cope with this issue, we trained a model on the entire training set (training and validation) for a few epochs, and manually looked at the data having DSC < 0.8.

By applying this approach we ended up removing 14 images from the training set before training our final model. Note that we used the same Dice threshold to look at predictions made on test data and IXI with our final model, which led to removing additional 12 images (7 test + 5 IXI). Note also that all images that were removed were clear FSL failures, not prediction failures, confirmed by visual inspection.

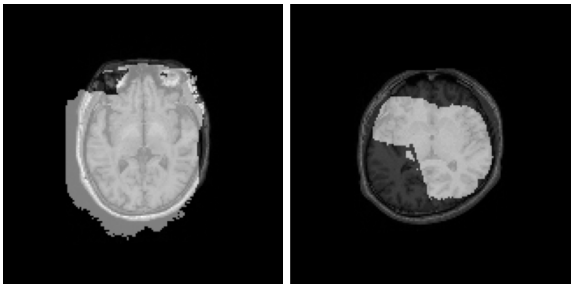


Figure 3: *Three instances of* `fsl_anat` *segmentation failure observed in our dataset.*

3 Results

Figure 4 depicts pair-wise 2D/3D comparative violin plots with jittering showing the distribution of the Dice coefficient for all MRI examinations across the collection of test data cohorts. From this, we observe close to negligible differences in performance between our 2D and 3D models. This is further illustrated by the results in Table 2, showing only small differences in the performance metrics.

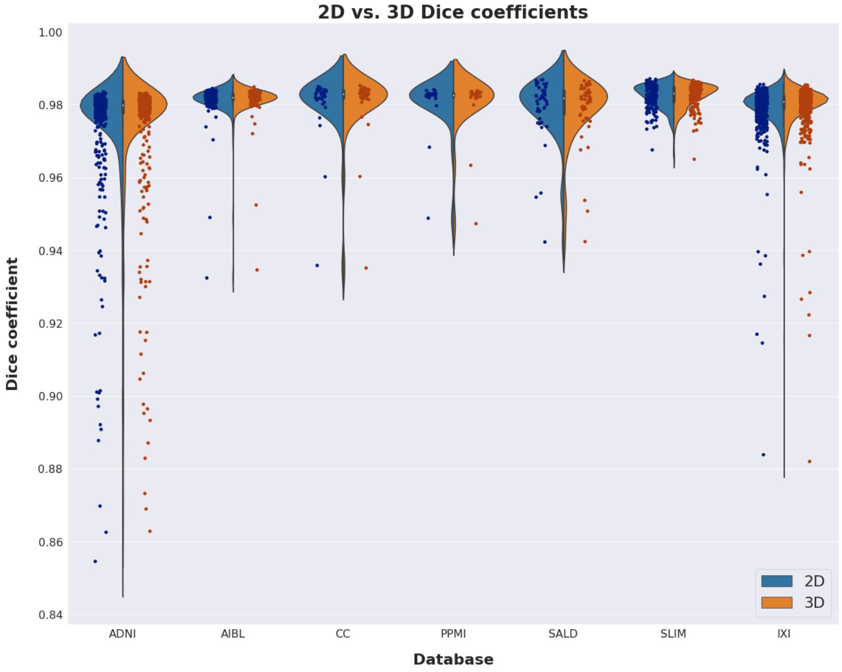


Figure 4: Violin plot of the Dice scores obtained by our models on the test dataset. Column names at the bottom of the plot refer to their database sources. The dots along the lower tail of the DSC distributions indicate outliers. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure4`.

	Test			**IXI**		
	Dice	**Jaccard**	**Hausdorff**	**Dice**	**Jaccard**	**Hausdorff**
2D U-Net	0.9778 (0.0131)	0.9569 (0.024)	5.6711 (4.7215)	0.9791 (0.0076)	0.9591 (0.0140)	5.7811 (5.4826)
3D U-Net	0.9781 (0.0133)	0.9574 (0.024)	5.0558 (6.4009)	0.9796 (0.0077)	0.9601 (0.0140)	5.9220 (2.7330)

Table 2: The average (SD) values of Dice score, Jaccard Index and Hausdorff distance on the test datasets (Test and IXI) for our 2D U-Net and 3D U-Net models.

On a standard CPU, making predictions, including loading the image data into memory, on the two test datasets (test and IXI) took 1.38 ± 0.05 s and 1.37 ± 0.02 s for the 3D model and 12.36 ± 0.57 s and 11.82 ± 0.54 s for the 2D model[5].

4 Discussion

Using a large collection of T1w MR images sourced from a variety of openly available datasets and a well-established set of FSL tools for automated generation of “ground truth" brain masks, we have constructed 2D and 3D models for fast and accurate skull stripping. On independent test sets our models were able to produce brain masks that are very close to those produced by the much slower FSL-based process ($\sim$ 10 mins per volume), and even in some cases demonstrating higher robustness than the slower approach (Fig. 5).

[5] On a single GPU the time for inference for the 3D model was 0.59 ± 0.05 s and 0.57 ± 0.01 s on the two test datasets

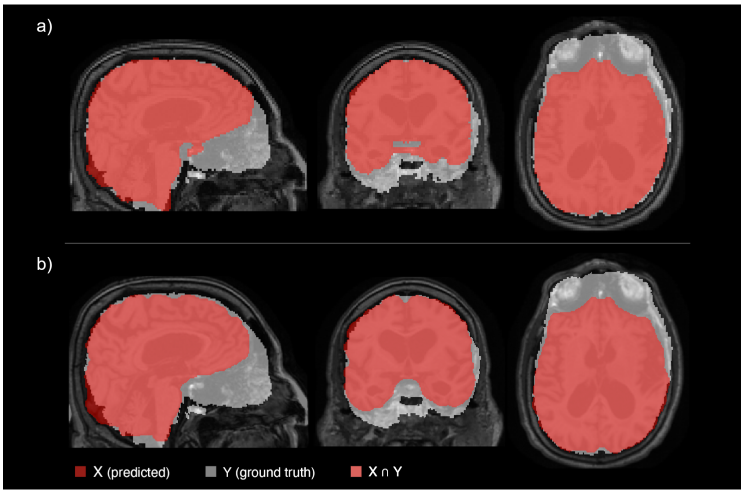


Figure 5: Comparison of (a) the brain mask produced by our 3D U-Net model on a dataset from ADNI achieving a poor Dice score, DSC < 0.9, and the corresponding ground truth FSL mask, and (b) the same comparison of the HD-BET model from [11]. Note the large amount of misclassification (extra-cerebral detection) of brain tissue by the "ground truth" FSL skull stripping procedure. This indicate that CNN-models may have some robustness advantages over FSL and perhaps also other traditional skull stripping . A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure5`.

In our comparisons between the 2D and 3D approaches we found similar performance as measured by Dice scores, Jaccard Index and Hausdorff distance, but also that the slice-by-slice based predictions necessary for the 2D approach made it significantly slower.

To decrease the variation in the training data images we performed bias-field correction before the images were fed to the network. This means that the networks have seen less bias than naturally occurs. To investigate the impact of this design decision we evaluated the trained models on the non-bias field corrected test images, reoriented to the standard MNI152 orientation, and also on and image with a high bias field shown in Fig. 2. Our 3D model had a Dice score of 0.978 ± 0.014 on the test set and 0.979 ± 0.008 on the IXI dataset when fed the uncorrected images. On the single high-bias field image displayed in Fig. 2, the model had a Dice score of 0.956 on the bias-field corrected image and a Dice score of 0.955 on the uncorrected image (Fig. 6).

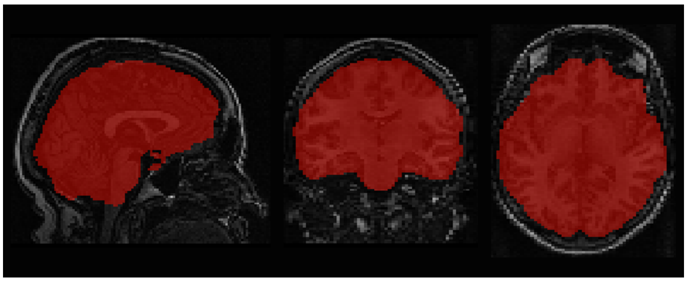


Figure 6: Predicition on a T1w image recorded at our own institution [5]. A color version of the image is available here: `https://tinyurl.com/skull-NIK2020-figure6`.

Having a fast and accurate skull stripping method can have practical utility as it can speed up larger image processing pipelines, e.g. for subcortical segmentation or segmentation of other regions of interest like brain tumors or lesions. Once the accuracy reaches a certain threshold, issues related to defining the ground truth becomes more important than increased accuracy at reproducing said ground truth labels. This can be illustrated by the different labels used in our work and in the HD-BET work of [11] described above. Feeding our test images through the trained HD-BET model results in an average Dice score of 0.9615 ± 0.0295. This does not mean that their model performs worse than ours at skull stripping, only that the ground truth labels used when training the models differs. Robustness also becomes more important than increasing the accuracy. As indicated in Fig. 5, CNN-based models may have an advantage here, but this requires further investigation.

Using our approach in a setting with various pathologies will require further investigations of its robustness to such variation in the images, and also clarification of “ground truth" consensus. Training the models on datasets that includes images with pathologies, and also adding an automized MRI quality control system based on e.g. MRIQC [23], would be natural next steps.

For thorough validation in a realistic setting, embedding the models in established workflows is key. At our hospital we have recently established a PACS, RIS and EDC system for research that integrates with the clinical systems. This enables real-world testing of this and other image processing methods, a crucial step for bringing deep learning research into practice [24].

5 Acknowledgments

Our work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

Data collection and sharing for this project was partly funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2.

[2] E. H. Shortliffe and B. G. Buchanan, “A model of inexact reasoning in medicine,” *Mathematical biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.

[3] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[4] T. Ching *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[5] M. A. Ystad, A. J. Lundervold, E. Wehling, T. Espeseth, H. Rootwelt, L. T. Westlye, M. Andersson, S. Adolfsdottir, J. T. Geitung, A. M. Fjell, I. Reinvang, and A. Lundervold, "Hippocampal volumes are important predictors for memory function in elderly women," *BMC Med Imaging*, vol. 9, no. 1, 2009.

[6] P. Kalavathi and V. S. Prasath, "Methods on skull stripping of MRI head scan images – a review," *Journal of Digital Imaging*, vol. 29, no. 3, pp. 365–379, 2016.

[7] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, 2002.

[8] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *NeuroImage*, vol. 45, no. 1, pp. S173–S186, 2009.

[9] B. Avants, A. Klein, N. Tustison, J. Woo, and J. C. Gee, "Evaluation of open-access, automated brain extraction methods on multi-site multi-disorder data," in *16th annual meeting for the Organization of Human Brain Mapping*, 2010.

[10] R. W. Cox, "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical research*, vol. 29, no. 3, pp. 162–173, 1996.

[11] F. Isensee *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[12] R. Dey and Y. Hong, "CompNet: Complementary segmentation network for brain MRI extraction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 628–636, Springer, 2018.

[13] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[14] G. Gavidia-Bovadilla, S. Kanaan-Izquierdo, M. Mataró-Serrat, A. Perera-Lluna, A. D. N. Initiative, *et al.*, "Early prediction of Alzheimer's disease using null longitudinal model-based classifiers," *PloS one*, vol. 12, no. 1, p. e0168011, 2017.

[15] K. A. Ellis *et al.*, "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease," *International psychogeriatrics*, vol. 21, no. 4, pp. 672–687, 2009.

[16] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Journal of neuroscience methods*, vol. 264, pp. 47–56, 2016.

[17] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[18] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, 2001.

[19] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, 2002.

[20] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[21] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv preprint arXiv:1707.02937*, 2017.

[22] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, International Society for Optics and Photonics, 2019.

[23] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites," *PloS one*, vol. 12, no. 9, 2017.

[24] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, 2020.

PULMONARY NODULE CLASSIFICATION IN LUNG CANCER FROM 3D THORACIC CT SCANS USING FASTAI AND MONAI

Kaliyugarasan, Satheshkumar, Lundervold, Arvid and Lundervold, Alexander Selvikvåg.

In International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), Volume 6, Number 7, 4 May 2021.

Special Issue on Current Trends in Intelligent Multimedia Processing Systems

Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using `fastai` and `MONAI`

Satheshkumar Kaliyugarasan[1,3]**, Arvid Lundervold[2,3], Alexander Selvikvåg Lundervold[1,3] ** *

[1] Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen (Norway)
[2] Dept. of Biomedicine, University of Bergen (Norway)
[3] Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen (Norway)

** These authors contributed equally to the current work

Received 9 November 2020 | Accepted 31 March 2021 | Published 4 May 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

We construct a convolutional neural network to classify pulmonary nodules as malignant or benign in the context of lung cancer. To construct and train our model, we use our novel extension of the `fastai` deep learning framework to 3D medical imaging tasks, combined with the `MONAI` deep learning library. We train and evaluate the model using a large, openly available data set of annotated thoracic CT scans. Our model achieves a nodule classification accuracy of 92.4% and a ROC AUC of 97% when compared to a "ground truth" based on multiple human raters subjective assessment of malignancy. We further evaluate our approach by predicting patient-level diagnoses of cancer, achieving a test set accuracy of 75%. This is higher than the 70% obtained by aggregating the human raters assessments. Class activation maps are applied to investigate the features used by our classifier, enabling a rudimentary level of explainability for what is otherwise close to "black box" predictions. As the classification of structures in chest CT scans is useful across a variety of diagnostic and prognostic tasks in radiology, our approach has broad applicability. As we aimed to construct a fully reproducible system that can be compared to new proposed methods and easily be adapted and extended, the full source code of our work is available at https://github.com/MMIV-ML/Lung-CT-fastai-2020.

Keywords

Convolutional Neural Networks, Fastai, Lung Cancer, Thoracic CT.

DOI: 10.9781/ijimai.2021.05.002

I. Introduction

Using convolutional neural networks is well-known to result in powerful tools to analyse medical images, across a variety of important applications [1], [2]. This approach to medical image analysis can lead to valuable insights and assistance in imaging diagnostics. The path from research to clinical practice is however slow and arduous, perhaps more so than is generally thought [2], [3]. But the number of software solutions on the market, with regulatory approval and aimed at diagnostic support, is growing, along with their adoption in hospital workflows.

In radiology, the computed tomography (CT) imaging modality is currently experiencing the highest impact of deep learning-based solutions. CT uses computer-processed combinations of many X-ray measurements taken from different angles to produce cross-sectional digital images (virtual slices) of specific regions or organs within the human body. This allows for non-invasive inspection of disease processes or lesions. Another prominent and widespread imaging modality is magnetic resonance imaging (MRI). It is based on quite different physical principles (nuclear spins in magnetic fields, spin excitation by application of radio-frequency pulses, magnetic resonance, and tissue specific and disease-related magnetization and relaxation phenomena) and enables exploitation of a large collection of measurement techniques and contrast mechanisms. Compared to CT, MRI examinations are generally more expensive, more time-consuming and less available. The signal properties are also more complex and typically multi-parametric, and proper interpretation puts high demands on radiologists' specialized training and experience. This partly explain why CT is more heavily used in daily routine radiology, and also why it is a popular target for the med-ical machine learning community [4].

Identifying and assessing structures in the lung from thoracic CT scans (chest CT) is a crucial task across multiple diseases involving the lungs and upper abdomen, e.g. lung cancer, chronic lung disease and pneumonia. Computer-aided diagnostic tools addressing chest CT is therefore an important area in medical imaging[1].

The diagnosis and follow-up of lung cancer patients using chest CT requires the identification of malignant tumors appearing as

* Corresponding author.

E-mail addresses: sathiesh.kumar.kaliyugarasan@hvl.no (S. Kaliyugarasan), arvid.lundervold@uib.no (A. Lundervold), allu@hvl.no (A.S. Lundervold).

[1] An area of particular relevance at the time of writing is the viral pneumonia caused by SARS-CoV-2 ([5], [6]).

pulmonary nodules (i.e. spots on the lungs). Distinguishing benign and malignant nodules is difficult, as the differences can be subtle and the malignancy potential is highly variable [7], but such assessment forms an important source of information for diagnosis and evaluation of progression and treatment responses. Indications of lung cancer can also appear as incidental findings on CT scans. As chest CT is widely used across a range of diseases and injuries, this represents an additional challenge for radiologists.

II. Related Work

Multiple studies have investigated how CNNs can be used in the context of lung cancer. Two recent and quite comprehensive reviews are [8], [9]. Below we highlight two illustrative examples of recent, related work.

In [10], the authors constructed an end-to-end system based on three 3D CNNs for the localization and categorization of lung cancer risk, using low-dose CT images as inputs. They achieved a test set ROC AUC of 94.4% using data from the National Lung Cancer Screening Trial (NLST), and a ROC AUC of 95.5% on an independent data set collected at Northwestern Medicine. A retrospective reader study was conducted, in which their model outperformed six experienced US board-certified radiologists. Their system had four main components: (i) a Mask R-CNN for instance segmentation used to produce lung segmentation masks; (ii) a 3D RetinaNet CNN trained to output ROIs around possible cancer lesions; (iii) a 3D version of Inception V1 trained to predict cancer diagnosis within one year directly from CT volumes; (iii) a CNN classifier trained on features extracted from the detected ROIs as well as features extracted from the volume model, outputting malignancy scores for each ROI. Their study was based on a combination of publicly available data from LUNA, LIDC and NLST, in combination with a large data set sourced from Northwestern Medicine that is not publicly available. The source code used in their work is not publicly available.

In [11], the authors construct *DeepLung*, a "cancer diagnosis system" based on two 3D CNNs that perform lung nodule detection and binary classification (benign vs. malign), respectively. For nodule detection they constructed a 3D Faster R-CNN with dual-path blocks and a similar encoder-decoder structure to the U-Net of [12], obtaining a FROC (Free Response Operating Characteristic) score of 84.2% on the LUNA16 data set [13] using a 10-fold patient-level cross-validation split. Their nodule classification model consisted of a 3D dual-path network extracting classification features, and a gradient boosting machine trained on the extracted features combined with raw nodule CT pixels and nodule size. They achieved a classification accuracy of 90.44%on the LIDC-IDRI data set using the same cross-validation approach as in LUNA16. The source code is available at https: //github.com/wentaozhu/DeepLung.

III. Main Contributions

Motivated by a lack of a common set of training data for machine learning models for lesion malignancy classification in the literature and what we see as important missing elements in how most CNNs for 3D medical imaging tasks are trained, our objectives are the following: (i) bring a set of techniques for training CNNs that have been shown to be highly impactful for 2D image classification to 3D by extending and incorporating ideas from the popular `fastai` library, and (ii) to provide a reproducible setup of data and model evaluation that can be used by other researchers aiming to train models to perform lung nodule classification. Our main contributions are:

1. We preprocessed and prepared the comparably large and well-annotated LIDC-IDRI data set (Section IV) for use in a binary malignancy prediction task, taking care to set aside a separate test set consisting of particularly well-characterized patients.
2. We constructed and trained a three-dimensional CNN using our novel extension to 3D of the `fastai` [14] deep learning library, combining it with features from `MONAI` (https://github.com/Project-MONAI/MONAI[2]), obtaining results comparable to the state-of-the-art in nodule classification and patient-level cancer diagnoses for the LIDC-IDRI data set.
3. We investigated the malignancy predictions by integrating a 3D version of gradient-weighted class activation mapping (Grad-CAM) [16] in our framework, enabling some element of *explainable AI* [17].
4. To ensure reproducibility and to ease further extensions or adaptions of our approach, we have made the source code openly available under a permissive open source license at https://github.com/MMIV-ML/Lung-CT-fastai-2020, in a tutorial-like Jupyter Notebook [18] that step through the process from data loading to result interpretations.

IV. Methods and Materials

A. Data Set

Using supervised learning with CNN models requires large amounts of labelled training data. For pulmonary nodule analysis, the data is typically obtained by manually labelling nodule locations and outlining lesions on CT images, a costly and hard to scale process hampered by intra- and inter-rater variability. Nevertheless, reasonably large annotated data sets with benign and malignant pulmonary nodules have been made openly available for researchers, reducing the entry price and increasing the pace of new research.

We used the Lung Image Database Consortium image collection (LIDC-IDRI), consisting of diagnostic and clinical lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [19][3]. The images were extracted from the picture archiving and communication systems (PACS) of seven different institutions and anonymized in accordance with HIPAA guidelines. The data collection was approved by the local IRBs of the seven participating LIDC-IDRI institutions. To each image there is associated the results of a two-stage anno-tation process involving four experienced thoracic radiologists. First, in a blinded-read phase, each radiologist independently reviewed the CT scans, marking lesions belonging to one of three categories (*nodule* ≥ *3 mm*, *nodule* < *3 mm*, and *non-nodule* ≥ *3 mm*), where the concept of "nodule" refers to a focal abnormality[4]. Then each radiologist (among a total of 12 radiologists coming from altogether five LIDC-IDRI institutions) assessed independently and subjectively each *nodule* ≥ *3 mm* for characteristics such as subtlety, internal structure, spiculation, lobulation, shape (sphericity), solidity, margin, and likelihood of malignancy. Each such nodule, having (by its size) a greater probability of malignancy than lesions in the other two categories, was marked regardless of presumed histology, e.g. a primary lung cancer, metastatic disease, a noncancerous process, or indeterminate in nature.

By design, reader consistency studies are not possible with the LIDC-IDRI data set as the order of the readers varies from instance to instance. However, the marks from up to four readers for a given

[2] Originally, we developed our extension of `fastai` and `MONAI` for 3D MRI of the head, as a tool for the estimation of brain age from MRI recordings (unpublished work and [15]) indicating our framework's general utility.

[3] See also https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[4] Some radiologists will argue that these three lesion categories could be somewhat artificial relative to clinical practice.

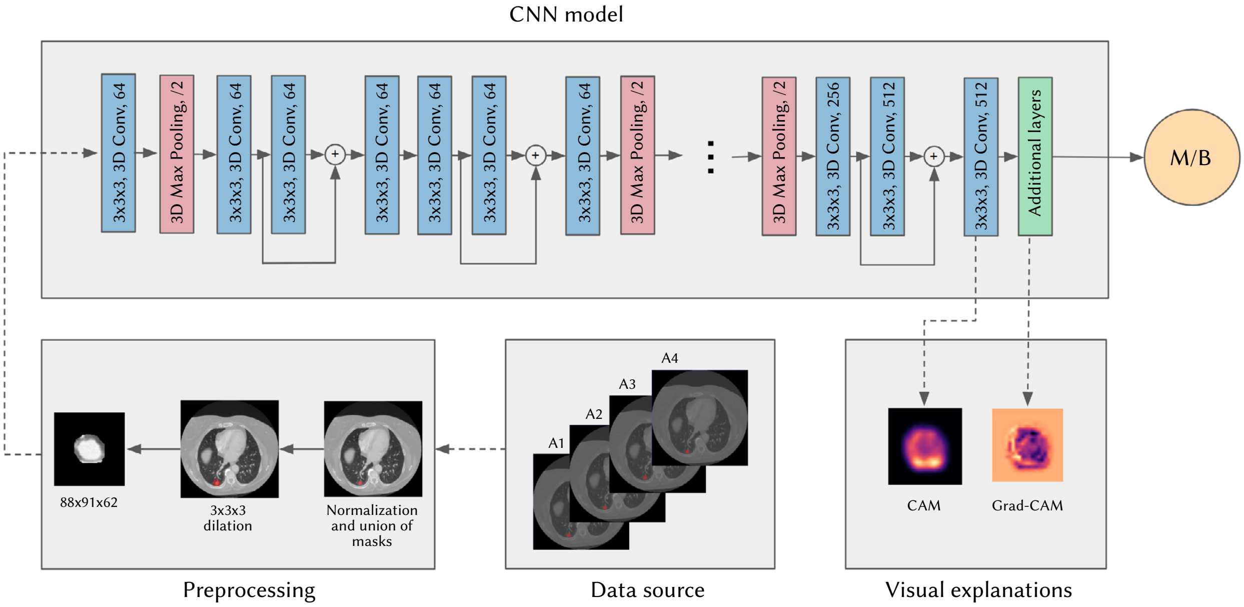


Fig. 1. The annotated images from our data source, LIDC-IDRI, are preprocessed by extracting 3D regions of interests around each of the nodules by taking the union of all the masks provided by expert annotations (e.g. A1–A4), before dilating the image slightly to capture some of the nodule surroundings. Using the expert assessment of malignancy, the resulting nodule images are used to train a 3D CNN model. This results in a nodule classification model with binary output: malignant (M) or benign (B). Our 3D implementation of class activation maps provides a visual explanation, here shown as a pair of 2D slices, indicating areas impacting our model's nodule classification decision. For further details, see the text and the accompanying code repository: https://github.com/MMIV-ML/Lung-CT-fastai-2020

lesion, using a five-point scale (a low score denoted likely benign nodule, a high score likely malignant), makes it possible to assess different degrees of reader agreement. Assessing inter-rater variability is very important to gauge the performance of systems aiming to automate the process. We therefore made an analysis of inter-rater variability regarding the "likelihood of malignancy" characteristic using the *Krippendorff's alpha* coefficient [20].

In our study, we have used a total of 2662 annotated nodules that were annotated as *nodule ≥ 3 mm* by at least one radiologists, collected from clinical thoracic CT scans of 1018 patients in the LIDC-IDRI data set.

B. Preprocessing

The voxels in a 3D CT recording are displayed in terms of relative radiodensity. More specifically, the signal intensities or attenuations in CT are expressed in Hounsfield units (HU). This is based on a linear transformation of the original attenuation coefficients in which the radiodensity of distilled water has HU = 0 and the radiodensity of air is set to HU = −1000. According to this HU scale, lung parenchyma is in the range [−700, −600], fat is [−120, −90], lymph nodes [+10, +20], and blood [+13, +50], to mention a few relative tissue types. In our CT data we considered voxels within a HU-range of [−1200, +600], and voxel values were normalized to the interval [0, 1] according to the transformation $x'' \mapsto x'' : x' = (x + 1200)/(1200 + 600);\ \ x'' = 0$ if $x' < 0, x'' = 1$ if $x' >… 1$, else $x'' = x'$.

For each CT scan of a subject, we collected all the radiologists segmentation masks. To ensure that we captured entire nodules we took the union of the masks. To make some of the surrounding context of each nodule available for the classification model, we dilated the resulting mask by adding 3 voxels to its boundary. The data set used to construct and evaluate our models was the constructed by applying the masks to the corresponding normalized CT and cropping to a cube containing the nodules. This gave us a total of 2662 3D images containing nodules. See Fig. 1 for an illustration of the preprocessing process.

We extracted each of the radiologists' subjective assessments of malignancy likelihood and computed the median scores across the readers for each nodule. If the median score for a nodule was < 3 we marked it as *benign*, if > … 3 as *malignant*. The nodules with median score 3 (indeterminant) were dropped from our data set. This gave us a total of 1106 benign nodules and 525 malignant.

C. Our `fastai` Extension and the 3D CNN Architecture

Our work is based on a combination of the `MONAI` deep learning framework and our own extension of the powerful `fastai` library built on top of PyTorch [14]. We have added functionality to support the construction, training and evaluation of three-dimensional convolutional neural networks, tailored for medical imaging-specific problems and file formats. In short, we have extended `fastai` to support 2D and 3D MRI and CT images by constructing new data loaders and data augmentation capabilities, and enabled the use of custom 3D CNNs while still supporting the highly impactful training techniques of `fastai`. This includes the learning rate finder [21] to find the optimum learning rate and the one-cycle learning rate policy (i.e. specific learning rate changes during the training, related to the concept of super-convergence [22], [23]).

The architecture of our 3D CNN is shown in Fig. 1. Each convolutional layer in our network consists of $3 \times 3 \times 3$ convolutions, followed by a batch normalization layer [24] and a rectified linear unit (ReLU) layer [25]. We add residual connections after each second convolutional layer. Each down-sampling block has a two-stride $2 \times 2 \times 2$ max-pooling layer.

To enable *discriminative learning rates*, i.e. different learning rates for different parts of the network, we divide the network into two layer groups: convolutional layers and additional layers. This also allow us to do gradual unfreezing, and eases the potential re-use of trained weights from the early layers for other tasks (i.e. *transfer learning*).

D. Training and Evaulation

To evaluate and get a robust estimate of our model's performance, we selected all the subjects in the LIDC-IDRI data set that have corresponding patient-level diagnoses as our test set (99 subjects, 238 nodules). The remaining data were divided into a training set (526 subjects, 1140 nodules) and a validation set (90 subjects, 255 nodules), using stratified sampling and no patient overlap between the sets. In order to deal with imbalanced classes in the training set (802 benign, 338 malignant), we over-sampled the malignant class by duplicating each sample.

Before feeding the images into the network, each image was padded to have the same volume dimension as the largest volume data × a scaling factor. We used data parallelism to train our model on four NVIDIA Tesla V100 32GB GPUs. Our training process was composed of two phases:

- Training a model on $44 \times 46 \times 31$ volumes, with weights randomly initialized (He initialization [26]).
- Training a final model on $88 \times 91 \times 62$ volumes, with weights initialized by copying the weights of the previous model.

This approach is known as progressive image resizing [27], a technique used to both reduce training time and to increase model performance. In our case, we found that it improved the accuracy on the validation set by almost two percentage points.

Our model was trained end-to-end in mixed precision [28] using the Adam optimizer [29]. The base value for the cyclic learning rate in the final model was set to 6×10^{-4} for frozen layers and 5×10^{-5} after unfreezing the layers, with learning rates for earlier layers scaled down by a factor of 20. We trained the model using a batch size of 128. For data augmentation we used random scaling with a factor from 1.0 to 1.1 and random rotation by an angle in the range [-35, 35]. As the geometry of the nodules can contain information about their malignancy, we only used shape-preserving morphisms. For regularization, we used a weight decay rate of 0.01 and a dropout ratio of 0.4, selected based on the performance on the validation data. Our final model was trained on the combined training and validation data for a few epochs, with a small cyclic learning rate, to also make use of the information contained in the validation data and its labels during model training.

E. Explainable AI and Class Activation Maps

As deep learning models are highly complex hierarchical objects with enormous amounts of parameters, there is an inherent "black-boxiness" to them. As they are increasingly being implemented across the medical imaging and decision making domains, this raises both technical challenges (how to open the black box?) and ethical conundrums (when is it OK to use predictions you cannot fully understand?). Using our extension of `fastai` we can produce what are called class activation maps (CAM) [30] and gradient-weighted class activation mapping (Grad-CAM) [16]. These are heat maps that can be used to indicate the importance of regions of an image for the model's classification, providing a relatively simple way to gain some explainability for image classification models, and potentially also to gain useful insights into the data used to construct the model.

CAM generates heat maps from the adaptive pooling layer, where the average of each cell across every channel is calculated. On the other hand, Grad-CAM uses the gradient information flowing into the last convolutional layer to produce heat maps, making it applicable to any CNN architecture.

A problem with these methods is that the resolution of the heat maps are the same size as the final convolutional layer. This means that we have to upsample them to the same size as the input images to highlight class-specific image regions. To mitigate this problem one can remove the pooling layers, but this will require more computational power due to larger spatial dimensions. In addition, overfitting is more likely to occur, which might reduce the performance of the network.

V. Experimental Results

Our test set consisted of 238 nodules from 99 subjects, 146 benign and 92 malignant. There were no overlap among train and test subjects. In addition to predicting nodule malignancy, we further investigated the models predictive capabilities by using the ground truth labels of patient diagnosis available in the LIDC-IDRI data set. The 99 patients in our test set were all diagnosed as either *malignant* or having *benign or non-malignant disease*. If one or more nodules from a patient was predicted to be *malignant*, we predicted malignant, else *benign or non-malignant disease*.

The results are displayed in Table I, Fig. 2 and Fig. 3.

TABLE I. Performance Metrics of Our Binary Classifier Predicting Single Nodules (N=238) and Patient Cases (N=99) in the Test Data Set: Accuracy (ACC), Precision (PREC) and Recall (REC). For the Patient Predictions We Give Performance Values Separately for Those Obtained by Our Model (CNN) and for Those Obtained by the Median Radiologist Assessments (Rad)

Classification task						
Nodule classification (%)			Patient classification (%)			
ACC	PREC	REC	Source	ACC	PREC	REC
92.4	85.6	96.7	CNN	75	86.8	78.7
			Rad	70	88.1	69.3

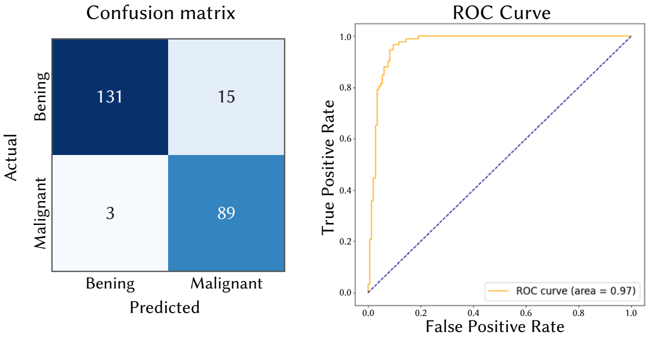
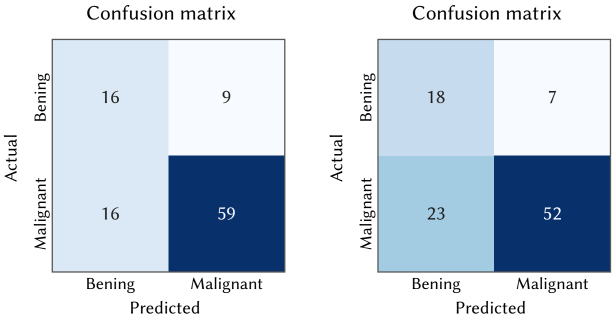


Fig. 2. Predicting the "likelihood of malignancy" in the test set of 238 nodules. (a) Confusion matrix. (b) ROC curve.



(a)

(b)

Fig. 3. Confusion matrices: (a) for the CNN predictions, (b) for the median malignancy scores by the radiologists. Note the additional cancer diagnoses captured by our CNN.

The mean score assigned to each nodule classified correctly as benign was 1.91 (SD 0.56) and as malignant 4.18 (SD 0.56). The nodules misclassified as benign had a mean score of 3.5 (SD 0.0) and those misclassified as malignant had a mean score of 2.23 (SD 0.4).

To assess the inter-rater variability and how the model compares to the human raters, we calculated the *Krippendorff's alpha* coefficient [20] for the 238 nodules. Krippendorff's alpha applies to any measurement level, can handle various number of raters and is invariant to the permutation and selective participation of raters. It also ignores missing data entirely. The independent and interchangeable rater panel per unit consisted of one to five radiologists using scores $s \in \{1$ (*most likely benign*), $2, \ldots, 5$ (*most likely malignant*)$\}$.[5] We note that the agreement on these subjective assessments were not very high. For the Krippendor's $\alpha \in [0, 1]$, $\alpha = 0$ is absence of agreement, and $\alpha = 1$ is perfect agreement. For the "likelihood of malignancy" we found Krippendorff's $\alpha = 0.49$, $\mathrm{CI}_{.025,.975} = [0.43, 0.54]$ (obtained by bootstrapping), indicating poor agreement among the raters.

The Krippendorff's alpha coefficient (in this case equivalent to Cohen's Kappa score) comparing the model's rating to the ground truth (determined by the median radiologist rating) was 0.84, $\mathrm{CI}_{.025,.975} = [0.78, 0.91]$.

The Krippendorff's alpha of the binary assessments of malignancy among the radiologists was $\alpha = 0.58$. By including the independent, CNN-based rater we obtained an increased alpha score to 0.68, indicating the usefulness of including this rater in the assessment of each nodule.

We applied our class-activation map approach described in Section IV.E to a selection of test nodules and CNN predictions. In general, getting better insight into CNN behavior and model predictions, both in cases where it classifies correctly and in cases where it fails, is of interest for several reasons. The class activation maps can provide discriminative information in image regions or part of the lesion being used by the model to predict the class label for the particular instance. This ability can at best introduce interpretability and trust in the model, or facilitate exploration and discovery of new features (image biomarkers) that might have a mechanistic relation to the disease process or disease state. In the present study, we did not fully explore the CAM approach or its potential by involving radiologists or pathologists, and the CAM results are anecdotal and not rigorously validated.

Some of the generated heat maps from our CNN model are presented in Fig. 4. By examining the malignant nodules (nodule 1 and nodule 2) and their corresponding heat maps, we can see that the lesion brims are highlighted, indicating that these regions are most important for the predictions. This might reflect typical malignant tumor growth characterized by central necrosis and viable tumor cells in a well-vascularized periphery. Another interesting finding was nodule 4, a nodule rated benign but classified as malignant by our model. This nodule was assessed by two radiologists deciding malignancy likelihood 2 and 3, respectively (i.e. towards benign), whereas the biopsy done on this nodule concluded that it was a malignant primary lung tumor.

VI. Discussion and Perspectives

We have addressed an important field of oncological radiology: the use of 3D CT scans to characterize focal lung lesions as benign or malignant. Using a large multi-center collection of well-organized CT examinations we constructed and trained a 3D CNN model to perform nodule malignancy classification.

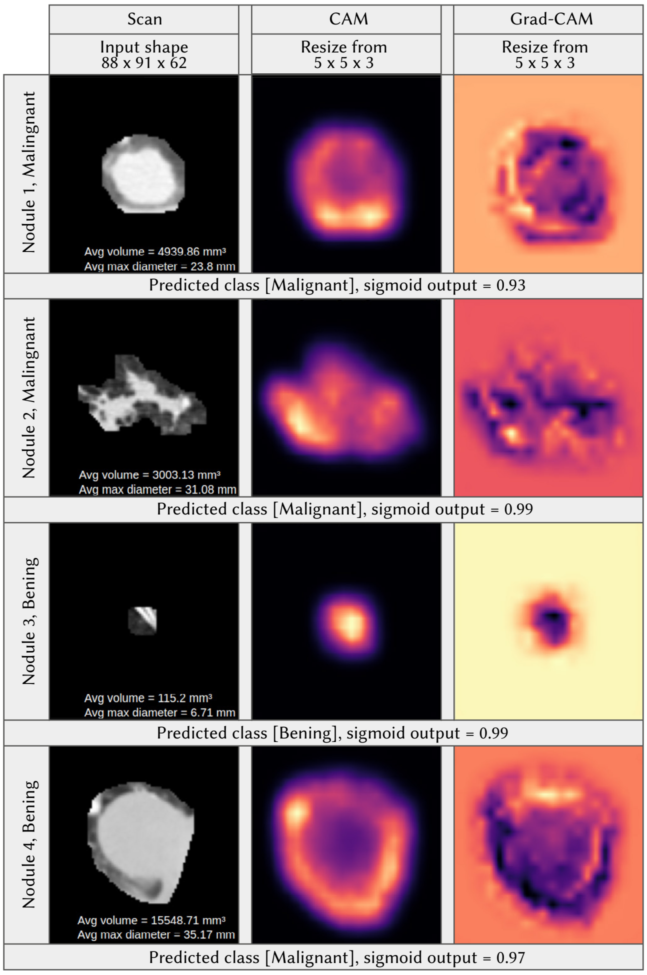

Scan
CAM
Grad-CAM
Input shape
88 x 91 x 62
Resize from
5 x 5 x 3
Resize from
5 x 5 x 3
Nodule 1, Malingnant
Avg volume = 4939.86 mm³
Avg max diameter = 23.8 mm
Predicted class [Malignant], sigmoid output = 0.93
Nodule 2, Malingnant
Avg volume = 3003.13 mm³
Avg max diameter = 31.08 mm
Predicted class [Malignant], sigmoid output = 0.99
Nodule 3, Bening
Avg volume = 115.2 mm³
Avg max diameter = 6.71 mm
Predicted class [Bening], sigmoid output = 0.99
Nodule 4, Bening
Avg volume = 15548.71 mm³
Avg max diameter = 35.17 mm
Predicted class [Malignant], sigmoid output = 0.97

Fig. 4. Examples of CAMs and Grad-CAMs for our model and the corresponding predictions and sigmoid outputs for the respective classes on a selection of four test set nodules.

Because CNNs automatically extract features from data, both interpretation and troubleshooting are more difficult compared to traditional machine learning models. For domains like medical diagnosis, where decision confidence is crucial, it is important to make sure that the results make sense. Otherwise, these models can easily end up performing worse than expected when used for real-world decision making. CAMs and Grad-CAMs generated from CNN models can be valuable for developers to gain some visual insights into models decision processes, helpful to identify data leakage, structural bias and for more comprehensive performance evaluation. In addition, the heat maps have the potential to detect local features that can be used as a biomarker for identifying malignant nodules. We implemented and explored these simple "explainable AI" techniques, assessing successful and unsuccessful nodule predictions.

Our model had a test set accuracy of 92.4% on the per-nodule malignancy classification task. On the patient-level malignancy classification task, our model had an accuracy of 75%. This gave an indication of the network's ability to pick up patterns corresponding to real nodule malignancy. As shown in Fig. 4, class activation maps can highlight regions of particular relevance for the nodule classifications, further indicating that the reasonableness of the features picked up by our CNN model.

In further work we will use the present system as a component in a detection + classification framework, obviating the need for manual annotation steps. We will test the system in the established

[5] The "likelihood of malignancy" characteristic is particularly subjective since the radiologists were not provided with any clinical information about the patients. As a general scaling guide, the likelihood of malignancy was rated under the assumption that the lesion was associated with a 60-year-old male smoker.

radiology research workflow at our hospital, through our "research PACS and RIS" system, enabling us to run arbitrary algorithms on locally recorded images. Such real-world testing is crucial to uncover and surmount the many technical obstacles faced when attempting to bring deep learning-based systems into practice [3]. Especially as it facilitates prospective investigations of the effect of combining the algorithm's predictions with radiologists' expertise, arguably the most interesting next step for research into applications of deep learning in medicine.

Acknowledgment

This work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

References

[1] A. S. Lundervold, A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[3] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, 2020.

[4] M. Brown, P. Browning, M. W. Wahi-Anwar, M. Murphy, J. Delgado, H. Greenspan, F. Abtin, S. Ghahremani, N. Yagh-mai, I. da Costa, *et al.*, "Integration of chest ct cad into the clinical workflow and impact on radiologist efficiency," *Academic radiology*, vol. 26, no. 5, pp. 626–631, 2019.

[5] C. Bao, X. Liu, Z. H., Y. Li, J. Liu, "Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis," *J Am Coll Radiol*, vol. Mar 25, pp. 1–9, 2020.

[6] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, P. Kanne, S. Raoof, N. W. Schluger, A. Volpi, J.-J. Yim, B. Martin, *et al.*, "The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society," *Chest*, vol. 158, no. 1, pp. 106–116, 2020.

[7] P. de Groot, B. Carter, G. F. Abbott, C. C. Wu, "Pitfalls in chest radiographic interpretation: blind spots," in *Seminars in roentgenology*, vol. 50, 2015, pp. 197–209, WB Saunders Ltd.

[8] D. Li, B. Mikela Vilmun, J. Frederik Carlsen, E. Albrecht-Beste, C. Ammitzbøl Lauridsen, M. Bachmann Nielsen, Lindskov Hansen, "The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: a systematic review," *Diagnostics*, vol. 9, no. 4, p. 207, 2019.

[9] A. Halder, D. Dey, A. K. Sadhu, "Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review," *Journal of Digital Imaging*, pp. 1–23, 2020.

[10] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[11] W. Zhu, C. Liu, W. Fan, X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 673–681, IEEE.

[12] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, Springer.

[13] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[14] J. Howard, S. Gugger, "fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[15] S. Kaliyugarasan, A. Lundervold, A. Lundervold, *et al.*, "Brain age versus chronological age: A large scale mri and deep learning investigation," 2020, European Congress of Radiology-ECR 2020.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[18] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter Notebooks — a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, p. 87, 2016.

[19] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[20] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[21] L. N. Smith, "No more pesky learning rate guessing games," *CoRR, abs/1506.01186*, vol. 5, 2015.

[22] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[23] L. N. Smith, N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 2019, p. 1100612, International Society for Optics and Photonics.

[24] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[25] V. Nair, G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[26] K. He, X. Zhang, S. Ren, J. Sun, "Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[28] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[29] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

Satheshkumar Kaliyugarasan

Satheshkumar Kaliyugarasan is a doctoral researcher at the Mohn Medical Imaging and Visualization Center focusing on machine learning in radiological imaging. In 2019 he completed his MSc degree in soft-ware engineering at the Western Norway University of Applied Sciences, Norway.

Arvid Lundervold

Arvid Lundervold is a professor of medical information technology at the University of Bergen and head of the Neuroinformatics and Image Analysis Laboratory in the Neural Networks Research Group, and co-leader of the Computational Medical Imaging and Machine Learning Group at the MMIV center. His research interests are image processing and pattern recogni tion, functional imaging, image registration, quantification and visualization, and mathematical modeling. Lundervold received an MD from the University of Oslo and a PhD in physiology from the University of Bergen.

Alexander S. Lundervold

A.S. Lundervold has a PhD in mathematics from the University of Bergen, Norway. He's currently working as an associate professor at the Western Norway University of Applied Sciences, and as a senior data scientist at the Dept. of radiology, Haukeland University Hospital, Norway. He leads the Computational Medical Imaging and Machine Learning Group at MMIV, together with A.L. His expertise lies in medical data analysis, with a particular focus on medical image processing and applications of machine learning to medicine.

PAPER D

FULLY AUTOMATIC WHOLE-VOLUME TUMOR SEGMENTATION IN CERVICAL CANCER

Hodneland, Erlend, Kaliyugarasan, Satheshkumar, Wagner-Larsen, Kari Strøno, Lura, Njål, Andersen, Erling, Bartsch, Hauke, Smit, Noeska, Halle, Mari Kyllesø, Krakstad, Camilla, Lundervold, Alexander Selvikvåg and Haldorsen, Ingfrid Salvesen.

In Cancers, Volume 14, Number 10, 11 May, 2022.

cancers

MDPI

Article

Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer

Erlend Hodneland [1,2,*,†], Satheshkumar Kaliyugarasan [1,3,†], Kari Strøno Wagner-Larsen [1,4], Njål Lura [1,4], Erling Andersen [1,5], Hauke Bartsch [1,6], Noeska Smit [1,6], Mari Kyllesø Halle [7,8], Camilla Krakstad [7,8], Alexander Selvikvåg Lundervold [1,3] and Ingfrid Salvesen Haldorsen [1,4,*]

[1] Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, 5009 Bergen, Norway; skka@hvl.no (S.K.); kari.strono.wagner-larsen@helse-bergen.no (K.S.W.-L.); njal.gjerde.lura@helse-bergen.no (N.L.); erling.andersen@helse-bergen.no (E.A.); hauke.bartsch@helse-bergen.no (H.B.); noeska.smit@uib.no (N.S.); alexander.selvikvag.lundervold@hvl.no (A.S.L.)

[2] Department of Mathematics, University of Bergen, 5020 Bergen, Norway

[3] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway

[4] Section of Radiology, Department of Clinical Medicine, University of Bergen, 5021 Bergen, Norway

[5] Department of Clinical Engineering, Haukeland University Hospital, 5021 Bergen, Norway

[6] Department of Informatics, University of Bergen, 5020 Bergen, Norway

[7] Department of Obstetrics and Gynecology, Haukeland University Hospital, 5053 Bergen, Norway; mari.halle@uib.no (M.K.H.); camilla.krakstad@uib.no (C.K.)

[8] Centre for Cancer Biomarkers, Department of Clinical Science, University of Bergen, 5021 Bergen, Norway

* Correspondence: erlend.hodneland@uib.no (E.H.); ingfrid.haldorsen@uib.no (I.S.H.)

† These authors contributed equally to this work.

Simple Summary: Uterine cervical cancer (CC) is a leading cause of cancer-related deaths in women worldwide. Pelvic magnetic resonance imaging (MRI) allows the assessment of local tumor extent and guides the choice of primary treatment. MRI tumor segmentation enables whole-volume radiomic tumor profiling, which is potentially useful for prognostication and individualization of therapy in CC. Manual tumor segmentation is, however, labor intensive and thus not part of routine clinical workflow. In the current work, we trained a deep learning (DL) algorithm to automatically segment the primary tumor in CC patients. Although the achieved segmentation performance of the trained DL algorithm is slightly lower than that for human experts, it is still relatively good. This study suggests that automated MRI primary tumor segmentations by DL algorithms without any human interaction is possible in patients with CC.

Abstract: Uterine cervical cancer (CC) is the most common gynecologic malignancy worldwide. Whole-volume radiomic profiling from pelvic MRI may yield prognostic markers for tailoring treatment in CC. However, radiomic profiling relies on manual tumor segmentation which is unfeasible in the clinic. We present a fully automatic method for the 3D segmentation of primary CC lesions using state-of-the-art deep learning (DL) techniques. In 131 CC patients, the primary tumor was manually segmented on T2-weighted MRI by two radiologists (R1, R2). Patients were separated into a train/validation (n = 105) and a test- (n = 26) cohort. The segmentation performance of the DL algorithm compared with R1/R2 was assessed with Dice coefficients (DSCs) and Hausdorff distances (HDs) in the test cohort. The trained DL network retrieved whole-volume tumor segmentations yielding median DSCs of 0.60 and 0.58 for DL compared with R1 (DL-R1) and R2 (DL-R2), respectively, whereas DSC for R1-R2 was 0.78. Agreement for primary tumor volumes was excellent between raters (R1-R2: intraclass correlation coefficient (ICC) = 0.93), but lower for the DL algorithm and the raters (DL-R1: ICC = 0.43; DL-R2: ICC = 0.44). The developed DL algorithm enables the automated estimation of tumor size and primary CC tumor segmentation. However, segmentation agreement between raters is better than that between DL algorithm and raters.

check for updates

Citation: Hodneland, E.; Kaliyugarasan, S.; Wagner-Larsen, K.S.; Lura, N.; Andersen, E.; Bartsch, H.; Smit, N.; Halle, M.K.; Krakstad, C.; Lundervold, A.S.; et al. Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer. *Cancers* **2022**, *14*, 2372. https://doi.org/10.3390/cancers14102372

Academic Editors: Keyvan Farahani and Bardia Yousefi

Received: 25 March 2022
Accepted: 5 May 2022
Published: 11 May 2022

Keywords: cervical cancer; deep learning; tumor segmentation

1. Introduction

Uterine cervical cancer is one of the leading causes of cancer-related deaths in women, particularly in developing countries [1]. For local staging, magnetic resonance imaging (MRI) is the preferred imaging modality due to its high soft-tissue resolution, conspicuously depicting the tumor and its boundaries to the surrounding tissue. Routine diagnostic work-up at many centers includes multiparametric MRI with diffusion weighted imaging (DWI), allowing the assessment of local tumor extent and maximum tumor diameter.

MRI radiomic tumor profiling involves the extraction of quantitative imaging information from segmented tumor masks using mathematical descriptors [2]. Radiomic tumor profiles have been linked to clinical phenotypes and prognosis for several cancers. Currently, there is a growing body of literature suggesting that the radiomic profile in CC is associated with prognostic factors [3–5], and predicts therapeutic response [6] and outcome [7–9].

Accurate tumor segmentation is a critical step in radiomic profiling since the radiomic data is specifically extracted from the segmented tumor volumes. Manual tumor segmentation in 3D by experts is, however, very labor intensive, making it unfeasible in routine clinical practice. Thus, a seamless clinical integration of whole-volume radiomic tumor profiling requires the development of robust platforms for accurate automated tumor segmentation. Previous CC studies applying deep learning (DL) networks for automated primary tumor segmentation on MRI data report highly variable Dice scores (Dice scores: 0.44-0.93) between DL segmentation and tumor segmentation derived by radiologists [10–13]. Furthermore, poor reproducibility of certain radiomic parameters derived from automatic tumor segmentations have been reported [12].

Traditional methods for medical image segmentation have relied upon techniques such as thresholding, edge detection, region-growing, clustering, or they have been based on the evolution of partial differential equations. However, over the past decade, DL-based segmentation methods have been shown to outperform classical segmentation methods, and have become state-of-the-art for complex segmentation tasks [14–17]. A common approach is to use models based on the U-Net architecture [18], which is an encoder-decoder convolutional neural network (CNN). U-Net-based models have been successfully employed in a wide range of medical imaging applications, including multi-parametric MRI tumor segmentation. The segmentation algorithm applied in the present work is an enhanced residual U-Net model [19].

This study aimed to use state-of-the-art DL libraries for automated CC segmentation. By training the platform on multiparametric pelvic MRI data in patients diagnosed with uterine CC we aimed to evaluate a DL algorithm for automated primary tumor segmentation in CC.

2. Methods

2.1. MRI Acquisitions

A total of 135 uterine CC patients diagnosed during 2009–2017 who underwent pre-treatment pelvic MRI (including DWI) and had visible tumors at MRI when assessed by two radiologists (hereafter referred to as Rater 1 and 2) were included in this study. Two of the patients were excluded due to poor image quality, and two patients with very large primary tumors (>1000 mL) were excluded, since more than two patients would be needed to be able to train a robust model on very large tumors. Thus, a total of 131 CC patients comprised the final study cohort.

The MRI examinations consisted of T2-weighted sequences and DWI with either two, three or four b-values. Apparent diffusion coefficient (ADC) maps were generated from mono-exponential fits to the DWI, using vendor-provided software at the scanner. T2-

weighted images, high b-value images and ADC-maps were all available when the raters manually segmented whole-volume tumor masks on the T2-weighted images. The MRI examinations were performed at multiple hospitals using different MRI scanners and protocols (see Table 1 for details).

The three imaging channels (T2-weighted, high b-value and ADC maps) were subsequently used in separate data sets for training (n = 90), validation (n = 15) and testing (n = 26) of the hlDL segmentation network.

Table 1. Summary of MRI protocols used in the study cohort (n = 131). The MRI data were acquired using different protocols, field strength and vendors. T2-weighted and diffusion-weighted imaging (DWI) acquisition parameters are reported as median values. FA = flip angle; FOV = field of view; mm = millimeters; ms = milliseconds; NA = not available; n = Number of patients in each category in terms of field-strength and vendor; s = seconds; T2 = T2-weighted; T = Tesla; TE = echo time; TR = repetition time. * Available b-values are reported, but not all b-values were available after export of the image data from the scanner.

	Parameter	**Siemens 1.5T**	**GE 1.5T**	**Philips 1.5**	**Siemens 3T**	**Philips 3T**
T2	Pixel spacing [mm] (inplane)	(0.39, 0.39)	(0.35, 0.35)	(0.40, 0.40)	(0.52, 0.52)	(0.35, 0.35)
	Matrix (x, y)	(512, 512)	(512, 512)	(512, 512)	(384, 384)	(512, 512)
	FOV [mm] (x, y)	(180, 180)	(180, 180)	(205, 205)	(200, 200)	(180, 180)
	TR [ms]	4790	3157	5362	4610	4074
	TE [ms]	100	81	100	94	110
	FA [degrees]	150	160	90	148	90
	Slice thickness [mm]	3.00	3.00	3.00	3.00	2.50
	Number of averages	2	2	6	2	2
	Interslice gap [mm]	0.50	0.00	0.30	0.30	0.25
	Number of slices	25	30	26	24	35
DWI	Pixel spacing [mm] (x, y)	(1.56, 1.56)	(1.37, 1.37)	(1.46, 1.46)	(1.43, 1.43)	(0.80, 0.80)
	Matrix (x, y)	(144, 144)	(256, 256)	(256, 256)	(144, 144)	(352, 352)
	FOV [mm] (x, y)	(250, 250)	(350, 350)	(375, 375)	(200, 200)	(280, 280)
	TR [ms]	3200	4000	1716.30	5640	3280
	TE [ms]	82	52	69.18	63	85
	FA [degrees]	90	90	90	180	90
	Slice thickness [mm]	4.00	5.00	5.00	3.00	4.00
	Number of averages	10	2	3	2	2
	Interslice gap [mm]	0.60	0.50	1.00	0.40	0.40
	Number of slices	22	25	30	25	33
	b-values [s/mm^2]	[0/50, 800/1000]	NA *	[0, 1000]	[0/50, 800/1000]	NA *
N	Number of patients	51	9	27	27	9

2.2. Inclusion Criteria

This retrospective study was conducted under Institutional Review Board (IRB)-approved protocols (2015/2333/REK vest) with written informed consent from all patients at primary diagnosis. All patients were diagnosed and treated at Haukeland University Hospital, Bergen, Norway. A total of 131 patients with histologically verified uterine cervical cancer who underwent pretreatment MRI between 2009 and 2017 were included. The patients were selected from a larger CC patient cohort scanned during 2002–2017 based on the following inclusion criteria for imaging data: (i) visible tumor on pelvic MRI confirmed by both radiologists; (ii) axial/axial oblique (relative to the long axis of the cervix) T2-weighted images; and (iii) axial/axial oblique DWI. An overview of patient characteristics in the training/validation and test cohorts is provided in Table 2.

Table 2. Patient characteristics of the training/validation cohort (*n* = 105) and the test cohort (*n* = 26). The two patient cohorts have similar clinicopathological characteristics. [1] Mann–Whitney U test. [2] Pearson's chi-square test. [3] Fisher exact test. [4] *n* = 97 for training/validation cohort and *n* = 25 for the test cohort. * Adenosquamous, neuroendocrine, and undifferentiated carcinomas; FIGO = International Federation of Gynecology and Obstetrics; IQR = Interquartile range; w/o = with and without.

Variable	**Train (*n* = 90) and Validation (*n* = 15) Data**	**Test Data (*n* = 26)**	***p***
Age (yrs.)			0.73 [1]
Median (IQR)	48 (37–60)	49 (41–59)	
FIGO (2009) stage			0.21 [2]
I	52 (49%)	14 (54%)	
II	27 (26%)	6 (23%)	
III	18 (17%)	5 (19%)	
IV	8 (8%)	1 (4%)	
MRI-assessed maximum tumor size (cm)			0.24 [1]
Median (IQR)	4.6 (3.0–5.6)	3.9 (2.5–5.1)	
Primary treatment			0.21 [2]
Surgery only	26 (25%)	9 (34%)	
Surgery and adjuvant therapy	63 (60%)	15 (58%)	
Primary radiotherapy w/o chemotherapy	12 (11%)	2 (8%)	
Palliative treatment	4 (4%)	0	
Histologic subtype			0.19 [2]
Squamous cell carcinoma	82 (78%)	21 (81%)	
Adenocarcinoma	18 (17%)	3 (11%)	
Other *	5 (5%)	2 (8%)	
Histologic grade [4]			0.76 [3]
Low/medium	80 (82%)	22 (88%)	
High	17 (18%)	3 (12%)	

2.3. Manual Tumor Segmentation

We used the open-source software ITK-SNAP (v. 3.6.0; www.itksnap.org, accessed on 16 June 2020) [20] for manual 3D tumor segmentation. The primary uterine cervical tumor was manually segmented on T2-weighted images, using axial oblique (when available) or axial images. Segmentations were performed by one radiologist in 105 patients (Rater 1 [K.W.L.]: *n* = 58; Rater 2 [N.L.]: *n* = 47) or both radiologists in 26 patients (comprising the test cohort). Rater 1 (R1) and Rater 2 (R2) had 12 and 7 years of experience in reading pelvic MRI, respectively. The radiologists were blinded to clinicopathologic patient information but had the DWI images available to support placement of tumor segmentations. The extracted 3D image mask was exported in the Neuroimaging Informatics Technology Initiative (NIfTI) file format [21].

2.4. Major Processing Steps

A flow chart of major processing steps is illustrated in Figure 1. The train and validation cohort comprised 105 randomly chosen patients used for training and validation of the 3D U-Net. Within this cohort, 90/105 sets went into training by stratified sampling based on tumor volume. The remaining 15/105 went into the validation data set and were used for reporting validation parameters during training. For a total of three times during the development period, the training and validation cohorts were selected at random from the set of 105 patients in order to increase the robustness of the algorithm and to avoid over-optimistic or over-pessimistic test results. The test cohort comprised 26 patients who had primary tumors manually segmented by both raters, serving as an unbiased test set for the evaluation of segmentation performance and inter-rater agreement.

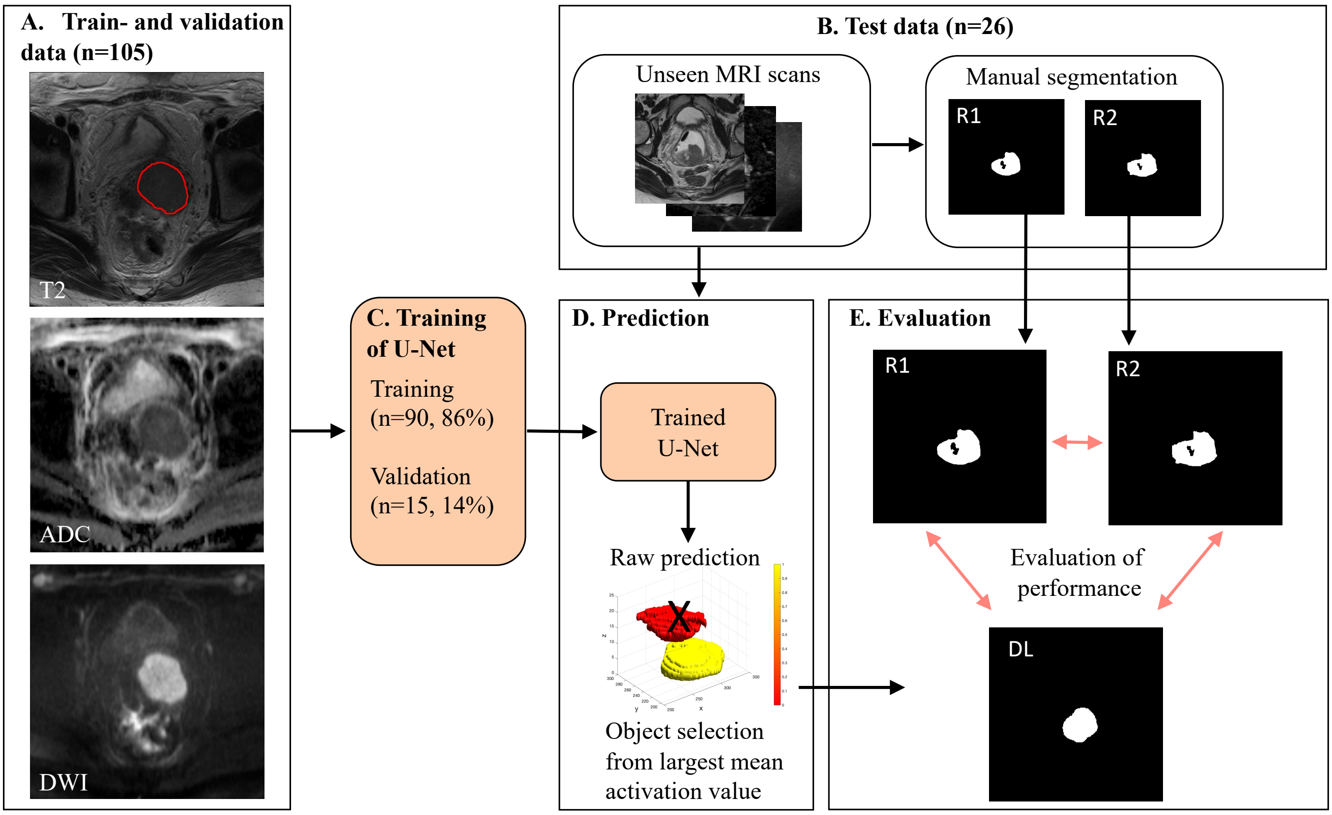


Figure 1. Graphical illustration of DL (deep learning) workflow and study setup. MRI data included T2-weighted images and diffusion weighted images (DWI), using high b-value images and apparent diffusion coefficient (ADC) maps at primary diagnostic work-up in 131 CC patients. (**A**) In the train and validation cohort (*n* = 105), primary tumor was segmented by one of the two expert raters (R1: *n* = 58, R2: *n* = 47). (**B**) The test cohort (*n* = 26), with primary tumor segmentations by both expert raters (R1 and R2), served as an unbiased test set for evaluating performance of the DL algorithm and inter-rater agreement. (**C**) The train and validation cohort (*n* = 105) was used to train a 3D U-Net using 90/105 (86%) cases for training and 15/105 (14%) cases for validation. (**D**) The trained network predicted raw tumor masks in the test data set (*n* = 26), identifying multiple regions in 23/26 cases. The object with the largest mean activation value was selected as primary tumor (yellow object). Other objects with lower mean activation values (red object) were removed from further analysis (indicated by a black cross). (**E**) DL-derived tumor masks were compared with manually segmented masks from R1 and R2, using Dice score and Hausdorff distances. R1 = Rater 1, R2 = Rater 2.

A sigmoid transformation was applied to the activation map compiled in the DL algorithm, providing a smooth function between zero and one. A final binary model prediction was derived by thresholding this function at a value of 0.5 [22]. However, thresholding leads to a binary map potentially containing multiple objects. In order to select the most probable mask object representing the primary tumor, we computed mean activation values within each individual object in the binary model prediction, with a neighborhood stencil size of 3. The object with the highest mean activation value was automatically chosen to represent the primary tumor. The activation map superimposed on a T2-weighted image of one patient is shown in Figure 2, generated using a 3D Slicer [23]. In this example, two potential tumor objects were identified. The object with the largest mean activation value was finally selected as a primary tumor. Segmentation performance was assessed in terms of DL-based tumor volumes and tumor masks' location compared with the R1 and R2 tumor segmentations.

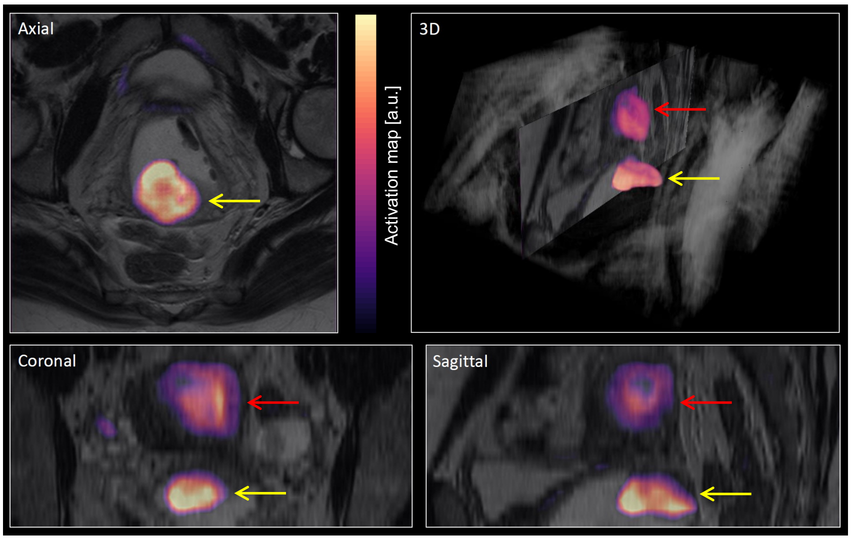


Figure 2. A visualization of the activation map (colored regions) from the DL (deep learning) segmentation superimposed on T2-weighted MRI (grayscale colormap) for three orthogonal planes and using 3D volume rendering. The activation map was later transformed with a sigmoid function and then thresholded, resulting in a binary prediction map. Two objects were identified in this patient: The object positioned in the uterine cervix (yellow arrows) had the largest mean activation value and was thus automatically selected to represent primary tumor. The object positioned in the uterine cavity/body (red arrows) had lower mean activation value and was thus excluded.

2.5. Evaluation of Segmentation Performance

To compare segmentation performance metrics, we used the Dice-Sørensen similarity coefficient (DSC) [24], measuring the degree of regional overlap between two segmentations. We also used the Hausdorff distance (HD) as a measure of maximum distance between segmented contours, as this is more sensitive to outliers in the segmentation shape, not sufficiently captured by DSC. The parameters are defined as

$$\mathrm{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}, \qquad \mathrm{HD}(x,y) = \max\left\{\delta(x,y), \delta(x,y)\right\}$$

where $|\cdot|$ is the cardinality and $\delta(x,y) := \sup_{x \in X} \inf_{y \in Y} d(x,y)$[25] for the Euclidean distance $d(x,y)$ between x and y. As a metric for segmentation performance, we also compare the estimated tumor volume between the hlDL algorithm and R1/R2.

The comparison of segmentation performance between R1 and R2 is referred to as inter-rater agreement. Similarly, the comparison of performance between the DL algorithm and R1 and R2 is referred to as DL-R1 and DL-R2, respectively. Median DSC and HD reported in Table 3 were adjusted to inter-rater agreement according to the formulas DSC(DL,R1/R2) ← DSC(DL,R1/R2) + (1-DSC(R1,R2) and HD(DL, R1/R2) ← HD(DL,R1/R2) − HD(R1,R2).

Patient characteristics for the training/validation and test data sets were compared using the Mann–Whitney U test for continuous variables and Pearson's chi-square test ($n > 5$ in any group) or Fisher exact test ($n \leq 5$ in any group) for categorical variables. Differences in median DSC and HD between DL and raters (DL-R1/R2), and between raters (R1-R2) were assessed using the Wilcoxon rank test. The agreement for primary tumor volumes between raters (R1-R2) and between DL and raters (DL-R1/R2) was reported using Bland–Altman plots and intraclass correlation coefficient (ICC). The difference in median tumor volume between raters and DL was assessed using Friedman's test. Correlations between tumor volumes and DSC or HD were tested using the Spearman correlation with H_0 of zero Spearman's ρ. Multiple linear regression was used to investigate statistical relations between T2- and DWI field-of-view (FOV) (defined as $\mathrm{FOV}_x \times \mathrm{FOV}_y$), T2- and DWI anisotropy (defined as slice thickness/max(pixel spacing x, pixel spacing y)), and field strength (1.5T or 3.0T) with DSC as the response variable. p-values below 0.05 are considered statistically significant. The statistical analyses were carried out in MATLAB using the Statistics and the Machine Learning Toolbox Version 12.0 (R2020b).

Table 3. Median (IQR = interquartile range) Dice score (DSC) and Hausdorff distance (HD) for tumor masks derived from DL (deep learning) segmentation compared to manual tumor segmentations by R1/R2. I: DL yields tumor masks with lower DSC and higher HD for DL-R1/R2 than that for R1-R2 (Wilcoxon rank sum, $p \leq 0.01$ and $p \leq 0.01$, respectively). II: Performance metrics of tumor masks after adjusting for median R1-R2 disagreement. The adjusted values yield higher DSCs and lower HDs for DL-R1/R2 when using DSC = 1 and HD = 0 as reference values for R1-R2 (ref. values). * Statistically significant; [1] Statistical testing and difference in estimates do not change from I to II; R1 = rater 1; R2 = rater 2.

	Measure	**Median Value of Estimate (IQR)**			**Absolute Difference (p-Value)**	
		***A.* (DL, R1)**	***B.* (DL, R2)**	***C.* (R1, R2)**	$\|A-C\|$ **(p)**	$\|B-C\|$ **(p)**
I. Unadjusted	DSC	0.60 (0.05, 0.78)	0.58 (0.09, 0.76)	0.78 (0.60, 0.83)	0.19 (0.01 *)	0.21 (0.005 *)
	HD [mm]	29.2 (14.5, 57.5)	30.2 (17.1, 55.9)	14.6 (9.80, 30.7)	14.6 (0.01 *)	15.5 (0.003 *)
II. Adjusted for R1-R2 disagreement	DSC	0.81	0.79	1 (ref.)	- [1]	- [1]
	HD [mm]	3.73	9.10	0 (ref.)	- [1]	- [1]

2.6. Implementation Details

Data sets were manipulated using the open-source, Python-based package `Imagedata` [26] for the reading and writing of image data between DICOM (https://dicomstandard.org, (accessed on 1 March 2018)) or NIfTI file format and `NumPy` arrays [27]. An in-house developed algorithm applying the geometric coordinate transformation specified within the DICOM image header was used for spatial alignment of the DWI data (ADC map and high b-value image) with the T2-weighted image using trilinear interpolation. After transformation, image voxel data for each patient was specified on the same spatial grid. Out-of-grid extrapolation values were set to zero.

We implemented our segmentation model using the 3D U-Net architecture from the MONAI framework (https://monai.io, [19], accessed on 17 December 2021), with 5 layers of 16, 32, 64, 128, 256 channels, respectively, each with downsampling and upsampling by a factor of 2, and a skip connection between them. The training was performed using our own extension of the fastai library [28,29], which simplifies training of three-dimensional convolutional neural networks using modern best-practices for training deep neural networks.

Before training, all MRI volumes were resampled to isotropic $(0.7 \times 0.7 \times 0.7)$ mm^3 voxel size using `RegularGridInterpolator` from `SciPy` [30]. The interpolation method was ‘trilinear’ for the MRI data, and ‘nearest neighbor’ for the binary mask data. They were channel-wise normalized using z-normalization (i.e., zero mean and unit standard deviation), and resized to 304 × 304 × 144 dimensions using either cropping or zero padding. The image data were of different matrix sizes, and the amount of cropping and padding was therefore different between data sets. We used data parallelism in PyTorch to train our model, a batch size of 4, and trained the model for 60 epochs using Dice loss function [31] on four NVIDIA Tesla V100 32 GB GPUs. We employed a Ranger optimizer [32] with an initial learning rate of 0.1, rapidly decreasing during the final few epochs using a cosine annealing scheduler, an idea that is related to the concept of super-convergence [33]. For data augmentation, we used random zooming by a factor in the range [1, 1.2], and random elastic deformations with 5 control points along each dimension of the coarse grid with a maximum displacement set f to 4 along each direction at each control point. The transformations were performed on the fly during training, with a probability set to 0.2 for each transformation. The weights of our final model were selected based on a callback that monitored the DSC on the validation data after each epoch, with the condition of saving the model if the performance of the validation data was improved by at least 0.005× DSC from the currently best model. Source code used in this work is openly available via GitHub (https://github.com/MMIV-DL/cervical-cancer-segmentation-2022, accessed on 21 March 2022).

3. Results

3.1. Train and Validation Metrics

Train and validation losses, as well as the DSC of the validation data set, are reported in Figure 3 as a function of epoch number. The train loss is steadily decreasing, indicating numerical stability of the optimization algorithm. Epoch number 55 (highlighted with bold in the figure) represented the optimal stopping point, when the validation DSC reaches a plateau and before it starts decreasing due to the effect of over-training. This approach minimizes the risk of over-training, potentially lowering general performance on unseen data sets. The Dice score in the validation data set reached a value of 0.52 when using this optimal stopping point (Epoch number 55) (Figure 3).

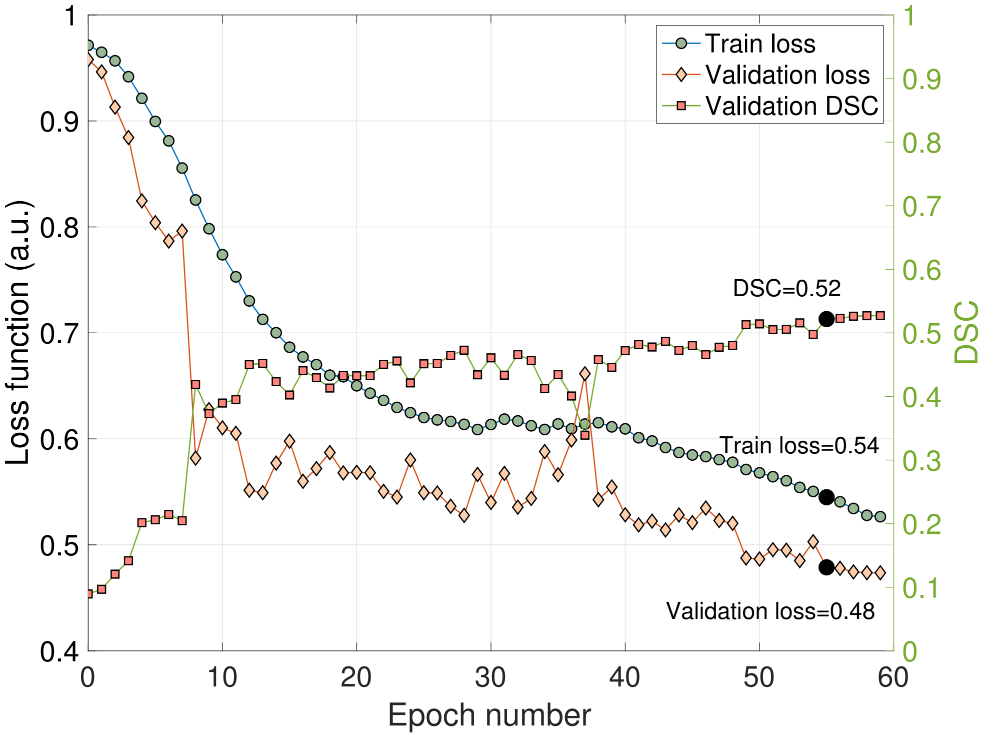


Figure 3. Train and validation losses (left axis) and Dice scores (DSC) (right axis) depicted as a function of epoch number. The train loss is smoothly decreasing, indicating numerical stability of the algorithm. The Dice score reaches a plateau, suggesting an optimal epoch number of 55 (black, solid dots). This epoch number yields optimal training performance of the network while minimizing the risk of over-training. a.u. = arbitrary units.

A histogram depicting the distribution of predicted objects in the test cohort (n = 26) is shown in Figure 4. The median (min, max) number of objects per patient was 7 (1.28). Only 12% (3/26) of the patients had one DL mask object, by definition representing the predicted primary tumor. For the vast majority (88%; 23/26), the predicted mask image contained multiple objects. For these patients, the object expressing the highest mean activation value was automatically selected to represent primary tumor.

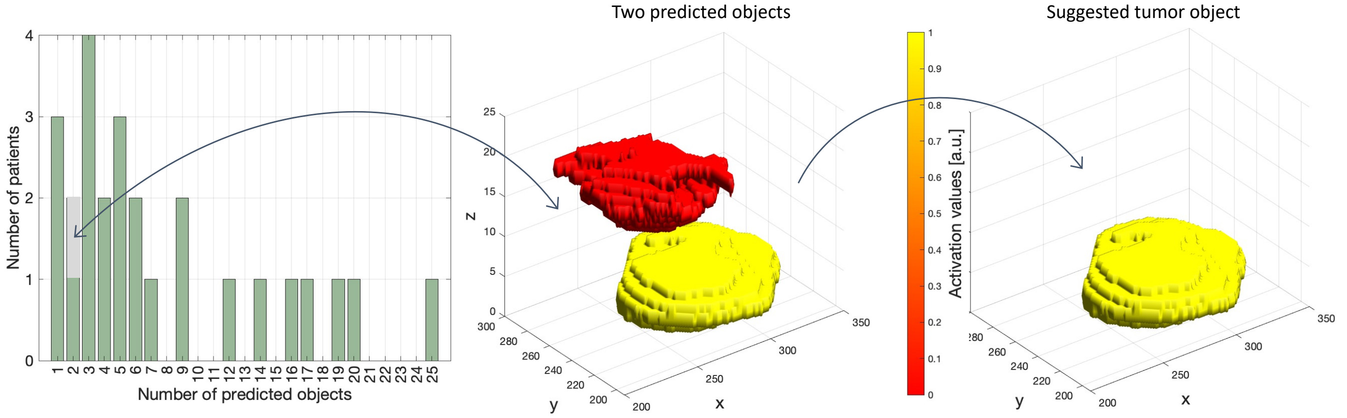


Figure 4. (**Left**): Histogram depicting number of objects in the prediction maps for the test cohort (n = 26) using a sigmoid-transformed activation map with a threshold of 0.5. In only 3/26 patients a single object was identified, whereas in 23/26 patients multiple mask objects were suggested. (**Middle**): Surface rendering depicting two objects (in red and yellow) in one of the patients having two predicted objects (grey box in histogram). The surface colors red/yellow indicate corresponding low/high mean activation values for the two objects (a.u. = arbitrarily units). (**Right**): In this patient with two suggested objects, the yellow mask with highest mean activation value was automatically selected as primary tumor.

3.2. Performance in Terms of DSC and HD

A summary of segmentation performance metrics in terms of DSC and HD for DL- and R1/R2 segmented primary tumor masks is given in Table 3. Segmentation performance of the DL algorithm is lower than that for the raters both in terms of median DSC (DL-R1: DSC = 0.60, DL-R2: DSC = 0.58, R1-R2: DSC = 0.78; Wilcoxon rank sum test, $p \leq 0.01$) and median HD (DL-R1: HD = 29.2 mm; DL-R2: HD = 30.2 mm, HD = 14.6 mm; Wilcoxon rank sum test, $p \leq 0.01$).

Box plots of DSC and HD for DL segmentation compared with that of R1 and R2 are depicted in Figure 5, reflecting the reported performance values in Table 3, I.

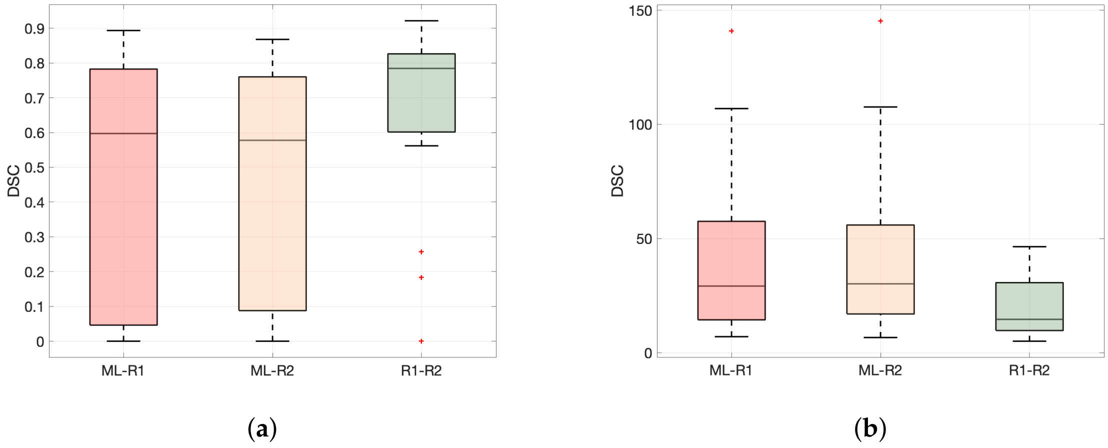
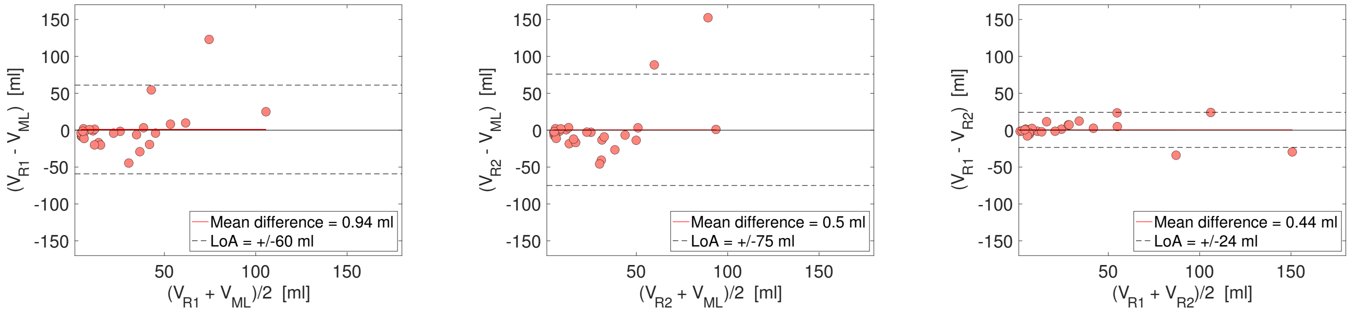


Figure 5. Comparison of (**a**) Median Dice coefficient (DSC) and (**b**) Median Hausdorff distance (HD) for segmentations by DL-R1, DL-R2, and R1-R2. Agreement between R1-R2 is significantly better than between DL and R1/R2 in terms of DSC and HD (Wilcoxon rank sum test, $p \leq 0.01$). The central line indicates the median, and the upper and edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers indicate the most extreme data points not considered to be outliers, while outliers are plotted individually using a '+' symbol. R1 = rater 1; R2 = rater 2.

After adjusting for R1-R2 disagreement, estimates of segmentation performance are significantly higher both in terms of median DSCs (DL-R1: DSC = 0.81, DL-R2: DSC = 0.79) and median HDs (DL-R1: HD = 3.73 mm, DL-R2: HD = 9.10 mm).

3.3. Performance in Terms of Reported Tumor Volume

Bland–Altman plots comparing reported primary tumor volumes based on segmentations by DL and R1/R2 and R1 and R2 are shown in Figure 6. The mean difference in tumor volume between DL-R1/R2 and R1-R2 was low for all comparisons (≤0.94 mL), suggesting high agreement in mean tumor volume. However, higher LoA of ±60/75 mL were found for DL-R1/R2 compared to R1-R2 with LoA of ±24 mL. There was no difference in median DL/R1/R2 tumor volumes (Friedman's test, $p = 0.10$). Agreement in terms of ICCs for log tumor volume for DL-R1 and DL-R2 was lower ($\text{ICC}_{\text{DL,R1}} = 0.43$ with 95% CI = (0.07, 0.70), $p = 0.01$, and $\text{ICC}_{\text{DL,R2}} = 0.44$ with 95% CI = (0.08, 0.70), $p = 0.01$, respectively) than that for R1-R2 ($\text{ICC}_{\text{R1,R2}} = 0.93$ with 95% CI = (0.85, 0.97), $p < 0.001$).



(**a**) (**b**) (**c**)

Figure 6. Bland–Altman plots comparing tumor volumes V [mL] from (**a**) DL (deep learning) and R1, (**b**) DL and R2 and (**c**) R1 and R2. Red lines indicate mean difference of the estimate, and dashed lines represent lower and upper limits-of-agreement (LoA). Mean difference in estimated tumor volumes is low for all comparisons, indicating a high agreement in mean primary tumor volume by all methods. However, LoA is higher for DL-R1/R2 than for R1-R2, indicating a higher individual disagreement for tumor measurements by DL-R1/R2 than by R1-R2. R1 = rater 1; R2 = rater 2.

A relatively weak but significant dependency of tumor volume on segmentation performance was observed for the DL algorithm (DL-R1: $\rho = 0.40$, $p = 0.046$; DL-R2: $\rho = 0.41$, $p = 0.039$, Spearman rank correlation) (Figure 7, upper row, left and middle panel). For R1-R2, large tumor size only tended to be positively correlated with segmentation performance (R1-R2: $\rho = 0.31$, $p = 0.12$; Spearman rank correlation) (Figure 7, upper row, right panel). All plots suggest a log-like relationship between increasing tumor size and DSC. We found no significant correlation between tumor volume and HD ($\rho \leq 0.24$, $p \geq 0.24$, Spearman rank correlation) (Figure 7, lower row). Patients with a low DSC < 0.2 and a small tumor volume < 50 mL are pairwise indicated in the upper and lower rows. For the ML-R1/R2 relation (Figure 7, left and middle columns), these cases ($n = 6$) had large HDs, whereas for the R1-R2 relation, similar cases ($n = 2$) had relatively low HDs (Figure 7, right column).

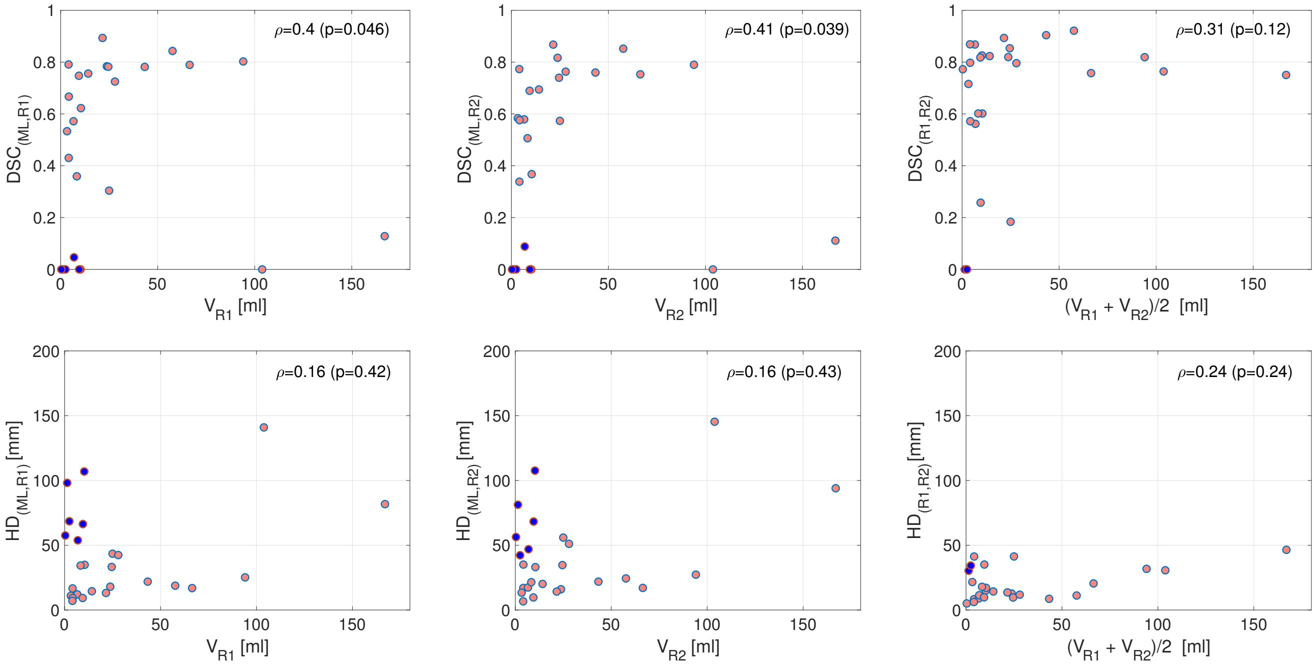


Figure 7. Tumor volume in relation to segmentation performance. (**Left**): R1 tumor volume. (**Middle**): R2 tumor volume. (**Right**): Mean tumor volume for R1- and R2 masks. (**Upper row**): Tumor volume against Dice coefficient (DSC). There is a weak but significant correlation between primary tumor volume and DSC for DL (deep learning)-R1/R2 (left and middle panel, $p \leq 0.046$). R1-R2 DSC only tended to be associated with tumor volume (right panel, $p = 0.12$). (**Lower row**): Plots of tumor volume against Hausdorff distance (HD). We found no significant correlation between primary tumor volume and HD for any of the associations DL-R1/R2 or R1-R2 ($p \geq 0.24$). Both rows: The same patients with (i) low DSC < 0.2 and (ii) a small tumor volume < 50 mL (estimated tumor volume for this condition is either R1 (left), R2 (middle), or mean (R1, R2) tumor volume) are simultaneously marked in blue in upper and lower panels, suggesting that patients experiencing a low DSC are normally high in HD for ML-R1/R2 (left and middle panels, the same $n = 6$ patients were identified). For R1-R2, patients with low DSC also have low HD (right panel, $n = 2$ patients). R1 = rater 1; R2 = rater 2; V = tumor volume, $\{\rho, p\}$ = Spearman rank correlation coefficient with associated p-value.

The multiple linear regression model reported in Table 4 revealed no linear relationship between field strength, T2/DWI anisotropy, and T2/DWI FOV as explanatory variables and DSC as response variable.

Table 4. Association between field strength, anisotropy T2 and DWI, field-of-view (FOV) T2 and DWI, and DSC using multiple linear regression. None for the MRI acquisition features had a statistical assocation to segmentation performance ($p \geq 0.33$, multiple linear regression).

	Estimate	**SE**	***p***
(Intercept)	0.08	0.36	0.82
Field strength	−0.06	0.15	0.71
Anisotropy T2	0.04	0.04	0.33
FOV T2	1.77	23.69	0.94
Anisotropy DWI	0.04	0.07	0.54
FOV DWI	4.10	6.11	0.51

4. Discussion

Patients diagnosed with uterine cervical cancer (CC) in high-income countries routinely undergo imaging by pelvic MRI, allowing the assessment of primary tumor extent and tumor invasion to surrounding tissue or pelvic lymph nodes. MRI-based whole-volume radiomic tumor profiling is promising for prognostication [3–5] and tailoring of cancer treatment [6–9]. However, the clinical utility of CC radiomic profiling is hampered by labor intensive manual tumor segmentations. In the current work based on 131 manually segmented primary CC lesions, we present a deep learning based algorithm for tumor segmentation yielding a fully automatic prediction of primary tumor position and boundaries.

This DL-based fully automatic approach for primary CC segmentations yielded relatively high segmentation performance (DL-R1: median DSC = 0.60, DL-R2: DSC = 0.58), although still lower than that for the expert raters (R1-R2: DSC = 0.78) (Table 3). Importantly, with a DSC of 0.78 for R1-R2, it is evident that substantial disagreement also exists when human experts define primary tumor boundaries in CC. Without using consensus segmentations across multiple raters, it seems inherently impossible to train a DL algorithm to yield better segmentation performance than that achieved for the raters involved in training the model. Thus, in an attempt to adjust for disagreement across raters we also report adjusted DSCs and HDs for the DL segmentation (Table 3). These adjusted performance metrics (DL-R1/R2: DSC = 0.81/0.79; HD = 3.73/9.10) are as expected better than the corresponding crude estimates (DL-R1/R2: DSC = 0.60/0.58; HD = 29.2/30.2) and may be considered as relatively good.

In the present study, the crude performance estimates are lower than that reported in some previous studies of CC tumor segmentation [10–13]. Bnouni et al. reported a DSC of 0.93 (using T2-weighed MRI) [13] (n = 15), Kano et al. reported a DSC score of 0.83 (using diffusion-weighted MRI) [11] (n = 98), and Lin et al. reported a DSC score of 0.82 (using multiparametric MRI) [12] (n = 169). However, these studies all used k-fold cross-validation applied to a train/validation data set for performance estimation and hyperparameter tuning. This setup is unfortunately not directly comparable to the present study since they did not estimate the performance of their DL algorithm in a separate and unbiased test data set [34].

Lin et al. presented a DL algorithm for automated tumor segmentations in CC using T2-weighted 3T MRI with DWI [12] (n = 169). Similar to our study, they used a separate test set to assess the performance of their DL algorithm, and report a DSC of 0.82. However, Lin et al. did not report inter-rater agreement as their manual tumor segmentations used for training of the DL network were by a single radiologist, however, with subsequent verification by a second radiologist. Thus, although the crude performance estimates of our DL algorithm (median DSCs of 0.60/0.58) seems inferior to that of the DL algorithm by Lin et al. (DSC of 0.82), the adjusted performance estimates for our DL algorithm (DSCs of 0.81/0.79) are quite comparable to that of their DL algorithm.

Interestingly, recent studies presenting DL algorithms for automated MRI tumor segmentations of other pelvic malignancies report performance metrics with DSCs in the range of 0.52–0.84 [35–39], i.e., prostate cancer (DSC of 0.52 using k-fold cross-validation [35] [n = 204]), endometrial cancer (DSC of 0.77/0.84 using a test set [36] [n = 139] and DSC of 0.81 using k-fold cross-validation [37] [n = 200]), and rectal cancer (DSC of 0.68/0.70 using a test set [38] [n = 140] and DSC of 0.70 using a test set [39] [n = 300]). Hence, our DSCs for the DL algorithm in CC (DL-R1: median DSC = 0.60, DL-R2: DSC = 0.58) are quite comparable to that of other pelvic malignancies. Similarly, inter-rater agreement in our study (R1-R2: DSC = 0.78) compares well with that reported in prostate cancer (DSC of 0.57, n = 78) [40] and rectal cancer (DSC of 0.83, n = 140) [38].

The use of one or multiple raters for the annotation of data sets may influence the segmentation performance of the DL algorithm. Interestingly, the study of Ji et al. [41] report higher performance of the DL algorithm when the algorithm utilizes the rich annotation information derived from manual segmentations from multiple raters. We included annotated data sets from two raters in the training data, however, with segmentations by a single rater for each patient in the training/validation set. Future work in CC segmentation should explore the value of rigorously incorporating information from multiple raters in order to maximize performance [41,42].

A further possible reason for variable segmentation performance for DL algorithms may be related to the extent of harmonization of input data. Although previous studies have identified dependencies of image resolution and noise characteristics on the reproducibility of radiomic features [43], we found no direct association between FOV, voxel anisotropy, and field strength on segmentation performance (cfr. Table 4). Still, the importance of using homogeneous imaging data in terms of standardized MR protocols for successfully training and applying a DL algorithm is not fully known. Despite z-normalization of the data prior to feeding the algorithm, variations in site, hardware and acquisition parameters in our study may have influenced the data in a way that has increased the complexity of the segmentation task. On the contrary, it is also possible that the algorithm becomes more robust by being exposed to variation in the training process, and that this may increase the performance of the algorithm when faced with new challenging segmentation tasks on images acquired at different sites and MRI scanners [44].

A majority of the raw predicted mask images contained multiple separate objects in 3D (23/26), with many of these being outside the uterine cervix. Although some of these additional objects could potentially represent extrauterine tumor tissue or metastases, they could not by definition represent primary tumor, and most of these objects were due to noise. A commonly used approach to handle multiple output regions is to threshold the predicted mask objects based on expected size [36,45], often in combination with various morphological operations [46]. We pursued an automatic approach that selected the most probable mask object based on maximum value of the average activation within the mask object. This approach attempts to take advantage of the inherent certainty built into the hlDL network, being expressed as high activation values whenever the network has high certainty for a tumor prediction, and oppositely expressing low activation values in the presence of low certainty.

The estimated tumor volume revealed no difference in median values between DL/R1/R2 (Friedman test, $p = 0.10$, and Bland–Altman plots, Figure 6). However, larger LoA for DL-R1/R2 compared to that of R1/R2 and higher ICC for R1-R2 (ICC = 0.93) than for DL-R1/R2 (ICC = 0.43/0.44) suggests that human experts reach a higher accuracy for tumor segmentation than the DL network. Future work must clarify how this observed difference in segmentation accuracy may influence radiomic feature extraction and potential prognostic modeling from corresponding radiomic signatures.

Interestingly, there was a positive, significant correlation between DSC and tumor volume for DL-R1/R2 ($\rho \geq 0.40$, $p \leq 0.046$) but only a tendency for R1-R2 ($\rho = 0.31$, $p = 0.12$) (Figure 7). This finding indicates that accurate tumor segmentation by the DL algorithm or even by human experts is easier to achieve if the tumors are relatively large. Our findings further support that whenever the DL method is failing in the presence of small tumors, the DL-suggested tumor mask is either very large or located far from the cervical lesion (Figure 7). Cases with inter-rater disagreement for small tumors had masks that were closer in space. This observed difference is probably because human experts have been trained to differentiate the uterine cervix from other organs.

Notably, similar findings with better segmentation accuracy in larger tumors have been reported in CC [12] ($n = 169$), and brain tumors [47] ($n = 69$). Identifying small tumors in CC can be a challenging task also for trained raters due to lack of contrast and inherent difficulties in distinguishing between normal and pathological tissue. Still, the weak relationship observed suggests that the challenges in retrieving accurate manual and automatic CC segmentations are only partly related to tumor size.

This study has several limitations. Our imaging data were acquired at different scanners with large variations in field-of-view, pixel size and field strength, leaving many images prone to substantial padding-/cropping effects and resizing in the preprocessing steps prior to feeding the network with data. The use of more standardized imaging protocols would potentially reduce the need for post-processing steps that are known to reduce data quality. However, it may be argued that this setup using imaging data derived from different scanners with their variable protocols, more truly reflects the standard imaging work-up that CC patients in general undergo. Furthermore, we excluded two patients with a tumor size > 1000 mL due to this small number being insufficient to train a model for large tumors. Thus, our findings in terms of performance in relation to tumor size may not be extrapolated to patients with extremely large tumors.

In conclusion, we have developed a DL algorithm for fully automatic primary tumor segmentation in CC that yields highly promising segmentation performance, although not yet reaching the same segmentation performance as human raters. With likely breakthroughs in DL technologies in the near future, this should motivate further development of similar DL platforms to enable automated radiomic tumor profiling in CC.

Author Contributions: Conceptualization, E.H., S.K. and I.S.H.; methodology, I.S.H., H.B. and E.H.; software, S.K., E.A. and A.S.L.; validation, E.H., S.K., K.S.W.-L., N.L. and I.S.H.; resources, I.S.H., C.K. and M.K.H.; writing—original draft preparation, E.H. and I.S.H.; writing—review and editing, all authors; visualization, E.H. and N.S.; supervision, I.S.H.; project administration, I.S.H. and C.K.; funding acquisition, I.S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study has received funding from The Western Norway Regional Health Authority (grant number 461120), Norwegian Research Council (grant number 311350) and Trond Mohn Foundation (grant number BFS2017TMT06 and BFS2016TMT01), Norway.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of REK vest (2015/2333/REK vest).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The Institutional Review Board approval did not allow for openly sharing of the MRI datasets. Source codes used for data processing and analyses are available in the Github repository https://github.com/MMIV-DL/cervical-cancer-segmentation-2022, accessed on 21 March 2022.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Varghese, B.A.; Cen, S.Y.; Hwang, D.H.; Duddalwar, V.A. Texture analysis of imaging: What radiologists need to know. *Am. J. Roentgenol.* **2019**, *212*, 520–528. [CrossRef] [PubMed]
3. Zhang, Q.; Yu, X.; Ouyang, H.; Zhang, J.; Chen, S.; Xie, L.; Zhao, X. Whole-tumor texture model based on diffusion kurtosis imaging for assessing cervical cancer: A preliminary study. *Eur. Radiol.* **2021**, *31*, 5576–5585. [CrossRef]
4. Xiao, M.; Ma, F.; Li, Y.; Li, Y.; Li, M.; Zhang, G.; Qiang, J. Multiparametric MRI-based radiomics nomogram for predicting lymph node metastasis in early-stage cervical cancer. *J. Magn. Reson. Imaging* **2020**, *52*, 885–896. [CrossRef] [PubMed]
5. Wang, T.; Gao, T.; Guo, H.; Wang, Y.; Zhou, X.; Tian, J.; Huang, L.; Zhang, M. Preoperative prediction of parametrial invasion in early-stage cervical cancer with MRI-based radiomics nomogram. *Eur. Radiol.* **2020**, *30*, 3585–3593. [CrossRef]
6. Sun, C.; Tian, X.; Liu, Z.; Li, W.; Li, P.; Chen, J.; Zhang, W.; Fang, Z.; Du, P.; Duan, H.; et al. Radiomic analysis for pretreatment prediction of response to neoadjuvant chemotherapy in locally advanced cervical cancer: A multicentre study. *EBioMedicine* **2019**, *46*, 160–169. [CrossRef]
7. Zhou, Y.; Gu, H.L.; Zhang, X.L.; Tian, Z.F.; Xu, X.Q.; Tang, W.W. Multiparametric magnetic resonance imaging-derived radiomics for the prediction of disease-free survival in early-stage squamous cervical cancer. *Eur. Radiol.* **2021**, *32*, 2540–2551. [CrossRef]
8. Lucia, F.; Visvikis, D.; Desseroit, M.C.; Miranda, O.; Malhaire, J.P.; Robin, P.; Pradier, O.; Hatt, M.; Schick, U. Prediction of outcome using pretreatment 18 F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* **2018**, *45*, 768–786. [CrossRef]
9. Lucia, F.; Visvikis, D.; Vallières, M.; Desseroit, M.C.; Miranda, O.; Robin, P.; Bonaffini, P.A.; Alfieri, J.; Masson, I.; Mervoyer, A.; et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 864–877. [CrossRef]

10. Torheim, T.; Malinen, E.; Hole, K.H.; Lund, K.V.; Indahl, U.G.; Lyng, H.; Kvaal, K.; Futsaether, C.M. Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning. *Acta Oncol.* **2017**, *56*, 806–812. [CrossRef]
11. Kano, Y.; Ikushima, H.; Sasaki, M.; Haga, A. Automatic contour segmentation of cervical cancer using artificial intelligence. *J. Radiat. Res.* **2021**, *62*, 934–944. [CrossRef] [PubMed]
12. Lin, Y.C.; Lin, C.H.; Lu, H.Y.; Chiang, H.J.; Wang, H.K.; Huang, Y.T.; Ng, S.H.; Hong, J.H.; Yen, T.C.; Lai, C.H.; et al. Deep learning for fully automated tumor segmentation and extraction of magnetic resonance radiomics features in cervical cancer. *Eur. Radiol.* **2020**, *30*, 1297–1305. [CrossRef] [PubMed]
13. Bnouni, N.; Rekik, I.; Rhim, M.S.; Amara, N.E.B. Context-Aware Synergetic Multiplex Network for Multi-organ Segmentation of Cervical Cancer MRI. In Proceedings of the International Workshop on Predictive Intelligence in Medicine, Lima, Peru, 8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–11.
14. Renard, F.; Guedria, S.; Palma, N.D.; Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* **2020**, *10*, 13724. [CrossRef] [PubMed]
15. Almeida, G.; Tavares, J.M.R. Deep learning in radiation oncology treatment planning for prostate cancer: A systematic review. *J. Med. Syst.* **2020**, *44*, 1–15. [CrossRef] [PubMed]
16. Lundervold, A.S.; Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Für Med. Phys.* **2019**, *29*, 102–127. [CrossRef]
17. Zhou, T.; Ruan, S.; Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* **2019**, *3*, 100004. [CrossRef]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
19. Kerfoot, E.; Clough, J.; Oksuz, I.; Lee, J.; King, A.P.; Schnabel, J.A. Left-ventricle quantification using residual U-Net. In Proceedings of the International Workshop on Statistical Atlases and Computational Models of the Heart, Granada, Spain, 16 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 371–380.
20. Yushkevich, P.A.; Piven, J.; Cody Hazlett, H.; Gimpel Smith, R.; Ho, S.; Gee, J.C.; Gerig, G. User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
21. Cox, R.; Ashburner, J.; Breman, H.; Fissell, K.; Haselgrove, C.; Holmes, C.; Lancaster, J.; Rex, D.; Smith, S.; Woodward, J.; et al. A (sort of) new image data format standard: NiFTI-1. Presented at the 10th Annual Meeting of the Organization for Human Brain Mapping, Budapest, Hungary, 13–17 June 2004.
22. Zhang, Y.; Chen, W.; Chen, Y.; Tang, X. A post-processing method to improve the white matter hyperintensity segmentation accuracy for randomly-initialized U-net. In Proceedings of the 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, 19–21 November 2018; pp. 1–5.
23. Kikinis, R.; Pieper, S.D.; Vosburgh, K.G. 3D Slicer: A platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative Imaging and Image-Guided Therapy*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 277–289.
24. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
25. Hausdorff, F. Grundzüge der Mengenlehre. In *SSVM*; Leipzig Viet: Leipzig, Germany, 1949.
26. Andersen, E. Imagedata: A Python library to handle medical image data in NumPy array subclass Series. *J. Open Source Softw.* 2022, *submitted*.
27. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
28. Howard, J.; Gugger, S. Fastai: A layered API for deep learning. *Information* **2020**, *11*, 108. [CrossRef]
29. Kaliyugarasan, S.K.; Lundervold, A.; Lundervold, A.S. Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using fastai and MONAI. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *6*, 83–89. [CrossRef]
30. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
31. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
32. Wright, L. Ranger—A Synergistic Optimizer. 2019. Available online: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer (accessed on 16 December 2021).
33. Smith, L.N.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*; International Society for Optics and Photonics: Bellingham, WA, USA, 2019; Volume 11006, p. 1100612.
34. Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

35. Lai, C.C.; Wang, H.K.; Wang, F.N.; Peng, Y.C.; Lin, T.P.; Peng, H.H.; Shen, S.H. Autosegmentation of Prostate Zones and Cancer Regions from Biparametric Magnetic Resonance Images by Using Deep-Learning-Based Neural Networks. *Sensors* **2021**, *21*, 2709. [CrossRef]
36. Hodneland, E.; Dybvik, J.A.; Wagner-Larsen, K.S.; Šoltészová, V.; Munthe-Kaas, A.Z.; Fasmer, K.E.; Krakstad, C.; Lundervold, A.; Lundervold, A.S.; Salvesen, Ø.; et al. Automated segmentation of endometrial cancer on MR images using deep learning. *Sci. Rep.* **2021**, *11*, 179. [CrossRef]
37. Kurata, Y.; Nishio, M.; Moribata, Y.; Kido, A.; Himoto, Y.; Otani, S.; Fujimoto, K.; Yakami, M.; Minamiguchi, S.; Mandai, M.; et al. Automatic segmentation of uterine endometrial cancer on multi-sequence MRI using a convolutional neural network. *Sci. Rep.* **2021**, *11*, 14440 . [CrossRef]
38. Trebeschi, S.; van Griethuysen, J.J.; Lambregts, D.M.; Lahaye, M.J.; Parmar, C.; Bakers, F.C.; Peters, N.H.; Beets-Tan, R.G.; Aerts, H.J. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci. Rep.* **2017**, *7*, 5301. [CrossRef]
39. Zhu, H.T.; Zhang, X.Y.; Shi, Y.J.; Li, X.T.; Sun, Y.S. Automatic segmentation of rectal tumor on diffusion-weighted images by deep learning with U-Net. *J. Appl. Clin. Med. Phys.* **2021**, *22*, 324–331. [CrossRef]
40. Liechti, M.R.; Muehlematter, U.J.; Schneider, A.F.; Eberli, D.; Rupp, N.J.; Hötker, A.M.; Donati, O.F.; Becker, A.S. Manual prostate cancer segmentation in MRI: Interreader agreement and volumetric correlation with transperineal template core needle biopsy. *Eur. Radiol.* **2020**, *30*, 4806–4815. [CrossRef]
41. Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; Zheng, Y. Learning calibrated medical image segmentation via multi-rater agreement modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12341–12351.
42. Warfield, S.K.; Zou, K.H.; Wells, W.M. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **2004**, *23*, 903–921. [CrossRef] [PubMed]
43. Roy, S.; Whitehead, T.D.; Quirk, J.D.; Salter, A.; Ademuyiwa, F.O.; Li, S.; An, H.; Shoghi, K.I. Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging. *EBioMedicine* **2020**, *59*, 102963. [CrossRef] [PubMed]
44. Bento, M.; Fantini, I.; Park, J.; Rittner, L.; Frayne, R. Deep Learning in Large and Multi-Site Structural Brain MR Imaging Datasets. *Front. Neuroinformatics* **2021**, *15*, 805669. [CrossRef] [PubMed]
45. Yu, W.; Fang, B.; Liu, Y.; Gao, M.; Zheng, S.; Wang, Y. Liver vessels segmentation based on 3D residual U-NET. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 250–254.
46. Tashk, A.; Herp, J.; Nadimi, E. Fully automatic polyp detection based on a novel U-Net architecture and morphological post-process. In Proceedings of the 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Athens, Greece, 8–10 December 2019; pp. 37–41.
47. Ngo, D.K.; Tran, M.T.; Kim, S.H.; Yang, H.J.; Lee, G.S. Multi-task learning for small brain tumor segmentation from MRI. *Appl. Sci.* **2020**, *10*, 7790. [CrossRef]

PAPER E

MULTI-CENTER CNN-BASED SPINE SEGMENTATION FROM T2W MRI USING SMALL AMOUNTS OF DATA

Kaliyugarasan, Satheshkumar, Dagestad, Magnhild H., Papalini, Evin I., Andersen, Erling, Zwart, John-Anker, Brisby, Helena, Hebelka, Hanna, Ansgar, Espeland, Lagerstrand, Kerstin M. and Lundervold, Alexander Selvikvåg.

To appear in the Proceedings of the of the 20th IEEE International Symposium on Biomedical Imaging (ISBI), 18-21 April 2023.

MULTI-CENTER CNN-BASED SPINE SEGMENTATION FROM T2W MRI USING SMALL AMOUNTS OF DATA

Satheshkumar Kaliyugarasan ⋆,1,2 *Magnhild H. Dagestad* 3,4 *Evin I. Papalini* 5,6
Erling Andersen 7 *John-Anker Zwart* 8,9 *Helena Brisby* 6,10 *Hanna Hebelka* 6,11
Ansgar Espeland 3,4 *Kerstin M. Lagerstrand* 5,6 *Alexander S. Lundervold* 1,2

[1] Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Norway
[2] MMIV, Dept. of Radiology, Haukeland University Hospital, Norway
[3] Dept. of Radiology, Haukeland University Hospital, Norway [4] Dept. of Clinical Medicine, University of Bergen, Norway
[5] Dept. of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Sweden
[6] Inst. of Clinical Sciences, The Sahlgrenska Academy at the University of Gothenburg, Sweden
[7] Dept. of Clinical Engineering, Haukeland University Hospital, Norway [8] Faculty of Medicine, University of Oslo
[9] Dept. of Research and Innovation, Division of Clinical Neuroscience, Oslo University Hospital
[10] Dept. of Orthopaedics, Sahlgrenska University Hospital, Sweden [11] Dept. of Radiology, Sahlgrenska University Hospital, Gothenburg, Sweden

ABSTRACT

Segmentation of the spinal tissues on MRI is the basis for quantitative analyses, but time-consuming if done manually. In this work, we construct a pipeline for automatic vertebrae segmentation from T2w MRI scans, assessing performance and generalizability by external validation. Our study used 15 scans from one site (Haukeland University Hospital, HUH) and 10 scans from another (Sahlgrenska University Hospital, SUH). MRI experts manually delineated the vertebral bodies Th12-L5 on all the HUH data and a subset of six scans from SUH. We trained multiple convolutional neural networks, assessing the performance in an experimental design tailored to small-data contexts and also on external data. Our best model achieved a mean Dice score of 0.899. This is comparable to results in the literature, but our system required much less training data. [1].

***Index Terms*—** Deep learning, image segmentation, MRI, lumbar and thoracic vertebrae

1. INTRODUCTION

Spine-related diseases have a massive impact on social costs and the health and quality of life for young and elderly people worldwide. The most common source of chronic disability for both sexes during the working years is low back pain. Magnetic resonance imaging (MRI) is an essential modality for clinicians to noninvasively assess the health status of the spine and visualize any pathology[1]. Some abnormalities in spinal tissues have been demonstrated to occur more frequently in patients with low back pain, including neoplastic, inflammatory, infectious, degenerative, and metabolic disorders. Therefore, objective evaluation of the vertebral tissue is warranted to monitor changes over time and enable comparisons between sites. Segmentation of the individual vertebrae is the basis for such quantitative analysis, but manual delineation is time-consuming. Hence, there is a need for automatic segmentation methods.

Deep learning models are extremely powerful for such automation tasks [2] but typically rely on humans to provide a large number of annotations. Deep learning approaches have been successfully applied to spine segmentation tasks using large annotated Computed Tomography (CT) datasets [3, 4]. For example, the multi-stage vertebra segmentation model for CT images by MONAI Label [5].

However, MRI is often the preferred modality for examining spinal diseases in clinical settings and is increasingly requested for patients with low back pain due to less radiation exposure and better soft-tissue visualization [1].

A few previous studies have investigated the use of deep learning for MRI-based vertebrae segmentation. Lu et al. [6] constructed ground truth annotations based on creating bounding boxes on the central slices of the sagittal T2w images (n=1000). The authors reported an average Dice score (DSC) of 0.93 (SD 0.02) using a 2D U-Net approach for the segmentation. In comparison, the delineated masks in our study are more detailed, making the segmentation task more complex. Furthermore, Lu et al. evaluated the performance on data from the same source as the training data, and they did not make their source code available. Zhou et al. [7] reported a mean DSC of 0.849 (SD 0.091) on their 2D U-Net approach

⋆ Corresponding author: skka@hvl.no

[1] The source code is available at `https://github.com/MMIV-ML/fastMONAI/tree/master/research`

using a dataset (n=57) with mid-sagittal slices derived from iterative decomposition of water and fat with echo asymmetric and least-squares estimation (IDEAL) spine MR images and annotated masks for L1-L5. The performance was evaluated on data from the same source as the training data. The trained model and the source code were made available at `https://github.com/zhoji/verteseg`. Recently, Gao et al. [8] trained a 2D U-Net model that achieved an average DSC of 0.882 (SD 0.018) based on T1w and T2w slices with manually annotated masks (n=40). The performance was evaluated on data from the same source as the training data, and the authors did not make their source code available. Lessman et al. [9] trained a 3D U-net-based CNN that achieved an average DSC of 0.944 (SD 0.033) using a three-fold cross-validation approach on an open lumbar spine MRI dataset [10] (n=23). The source code was not shared, making direct comparisons challenging[2].

In the present work, we train a 3D segmentation model in a setting with small amounts of data and evaluate the performance on external data. Furthermore, we share the complete source code to construct and train our models and the learned weights, enabling other researchers to produce segmentation masks for new T2w MR images.

Open in Colab [3]

2. METHODS AND MATERIALS

2.1. Data

We used 25 T2-weighted MRI scans of unique patients from Haukeland University Hospital (HUH) (n=15) and Sahlgrenska University Hospital (SUH) (n=10).

The data from HUH were part of the AIM-study (*Antibiotics In Modic changes*) [11], where all images were acquired on a Siemens MAGNETOM Avanto 1.5 T MRI scanner (slice thickness = 4 mm, FOV 300x300 mm^2, 384x269 matrix, interslice gap = 0.4 mm, repetition time = 3700 ms, echo time = 87 ms, number of acquisitions = 2) [12].

The data from SUH were acquired on a Siemens MAGNETOM Aera 1.5 T MRI scanner (slice thickness = 3.5 mm, FOV 300x300 mm^2, 384x384 matrix, interslice gap = 0.7 mm, repetition time = 3500 ms, echo time = 95 ms, number of acquisitions = 1) [13].

An MRI expert conducted the manual segmentation of the vertebrae with consensus from a senior radiologist using the open-source software ITK-SNAP [14]. The vertebrae were segmented on all slices in the image volume, visualizing the vertebral body. The annotations included the vertebral body bone marrow but not the vertebral cortex. Figure 1 shows an example of the annotated data.

To evaluate our model on an external dataset, we used the data from Chu et al. [10], consisting of 23 scans acquired at 1.5 T in sagittal orientation. Seven vertebrae are manually segmented on each image (Th11–L5). We manually removed Th11 to be consistent with the above data sets.

2.2. Methods

To construct our pipeline shown in Fig. 1, we used our open source fastMONAI library[4], used in previous studies [16, 17, 18]. fastMONAI is a low-code library built on top of fastai [19], MONAI [20], and TorchIO [21], making it easier to construct, use and train powerful deep-learning models for various medical imaging tasks.

The neural network architecture used in this study was an enhanced 3D U-Net[22] as implemented by the MONAI library[20], with layers of 16, 32, 64, 128, 256 channels, each with downsampling and upsampling by a factor of 2 and a residual connection between them. Based on inspection of the variation in the HUH training data, we decided to resample all volumes to 4.4 x 0.78 x 0.78 mm^3 voxel size, do z-normalization (i.e., zero mean and unit standard deviation), and resizing to 16 x 400 x 400 using zero-padding or cropping.

We set the number of epochs to 150 and used a batch size of 4. The models were trained using the Ranger optimizer [23, 19] with an initial learning rate of 0.01 that was decayed using flat-cosine-annealing where the last 25% of the training follows a cosine function as it slows down [24, 19]. We implemented Focal Tversky loss function, a generalized Focal loss function based on Tversky [25]. These choices were made based on experiments conducted in our cross-validation experiment (Experiment 1 below).

All our models were trained using data augmentation. We used data augmentation random gamma correction with log gamma value of -0.2 to 0.2, random scale factor from 0.9 to 1.1, random rotation $[-5°, 5°]$, and random elastic deformation with five control points and max displacement of 5.5 to simulate anatomical variations. The augmentations were applied on-the-fly during training.

Each model's training took approximately 25 minutes on an NVIDIA Titan RTX GPU, consuming roughly 4.5GB VRAM during training.

In a postprocessing step, we extracted three-dimensional connected components. We used them to remove small false positives (defined as $<= 20\%$ of the average vertebral bodies in the dataset) and calculate each vertebral body's volume.

See the accompanying source code for additional details.

[2] We did, however, use our methods in a three-fold cross-validation based on the same data as in [9] and achieved an average DSC of 0.94 (SD 0.01), indicating similar performance for the two approaches.

[3] `https://github.com/MMIV-ML/fastMONAI/tree/master/research`

[4] fastMONAI is available at `https://fastmonai.no`

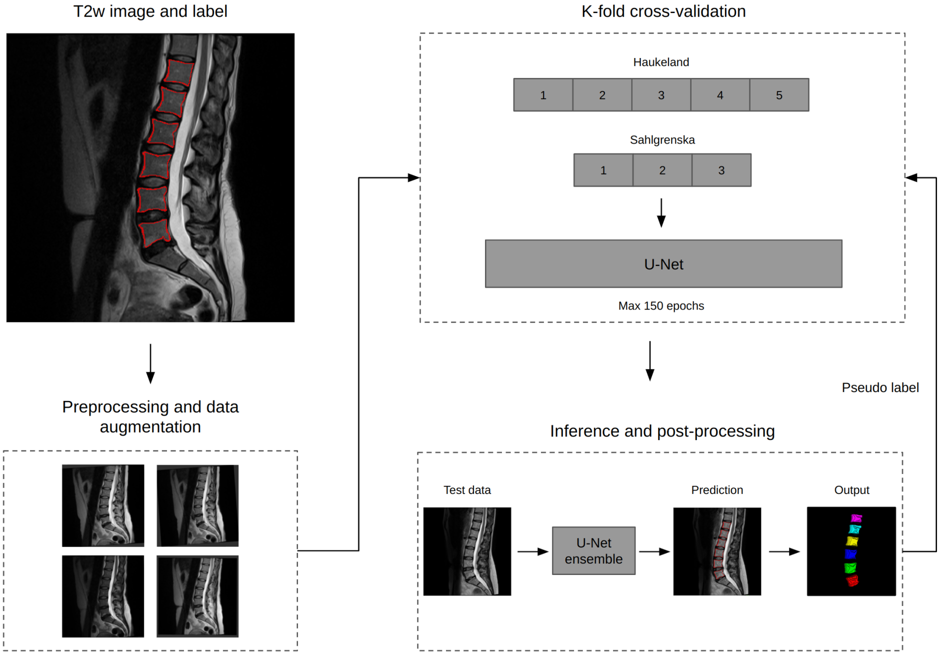


Fig. 1: The figure illustrates our proposed preprocessing, training, and evaluation pipeline. First, the T2w MRI scans and the corresponding labels were preprocessed by resampling, z-normalization, and zero-padding or cropping. Next, we expanded the training data set using data augmentation strategies and trained multiple U-Net-based segmentation models in two K-fold cross-validation setups. The results were evaluated on test data using the Dice coefficient and the resulting vertebrae volumes. In the first experiment, we used five-fold cross-validation based on only the data from HUH, evaluated the generalization ability on all the labeled SUH data. In the second experiment, we added data from SUH to the training set in a three-fold cross-validation setup, again evaluated the results using the Dice coefficients and the volumes. In our final set of experiments, we used an ensemble of the three models from the previous experiment to produce predicted labels on unlabeled data from SUH, first directly and then via a semi-supervised setup based on pseudo-labeling [15]. Finally, we evaluated the results using external data.

3. EXPERIMENTS AND RESULTS

We used K-fold cross-validation to estimate the generalizability of the models before applying them to unseen test data. For each validation image in a fold, we compared the predicted mask with the ground truth in terms of Sørensen-Dice similarity coefficient (DSC) and Hausdorff distance (HD). We performed the following experiments:

1. External validation, train on data from one site, assess performance on data from another
2. An approach for fine-tuning on target data
3. Semi-supervised learning using pseudo labeling
4. External evaluation on an open, annotated dataset [10]

The results are reported, and the experiments are explained further in Table 1 and in the Bland-Altman plots of Figure 2.

4. DISCUSSION

We presented an approach to vertebral body segmentation from T2w MRI scans using two independent sources. We designed and investigated a training setup tailored for contexts with limited data, evaluating our models' performance on external data. The average performance across the test set subjects indicated that it is possible to construct useful models from limited amounts of manually segmented vertebrae images, which is promising for future implementation in a clinical workflow. However, we observe some performance variability among the subjects. A further investigation into the sources of this variability–scanner settings, ground truth labeling, anatomical variation, etc., is warranted. Before such a system can be implemented, additional hurdles must be overcome. A natural next step is to incorporate the model in an active learning setup and use it to label a larger dataset (e.g., [26]). This will reduce the time needed for manual delineation while simultaneously training the model on more data. We plan to deploy such a setup in the research PACS infrastructure of our health region[5], where we can evaluate its clinical usefulness in a context that is familiar to radiologists.

[5] See `https://mmiv.no/wiml`

Experiment 1

Dataset	Cases	raw DSC	post-processed DSC	raw HD	post-processed HD
HUH	15	0.894 ± 0.034	0.896 ± 0.033	51.605 ± 51.737	45.629 ± 59.750
SUH	6	0.855 ± 0.036	0.859 ± 0.035	28.257 ± 15.002	19.221 ± 17.586

Experiment 2

SUH	6	0.886 ± 0.025	0.890 ± 0.025	31.504 ± 4.384	15.332 ± 12.927

Experiment 3

SUH	6	0.896 ± 0.034	0.898 ± 0.031	25.993 ± 25.889	16.329 ± 15.141

Experiment 4

Public dataset [10]	23	0.896 ± 0.021	0.899 ± 0.02	53.4 ± 39.733	12.052 ± 11.6195

Table 1: Experiment 1: models trained on HUH data using five-fold cross-validation and assessed on SUH data. Experiment 2: trained on all HUH data and some SUH data using three-fold cross-validation with SUH data in the validation folds. Experiment 3: an ensemble of the models from Experiment 2 was applied to the unlabeled data from SUH, and the generated labels for the unlabeled data were used as pseudo labels in the same three-fold cross-validation training process as above. Experiment 4: the ensemble from experiment 2 was used to produce labels for the unseen dataset provided by [10].

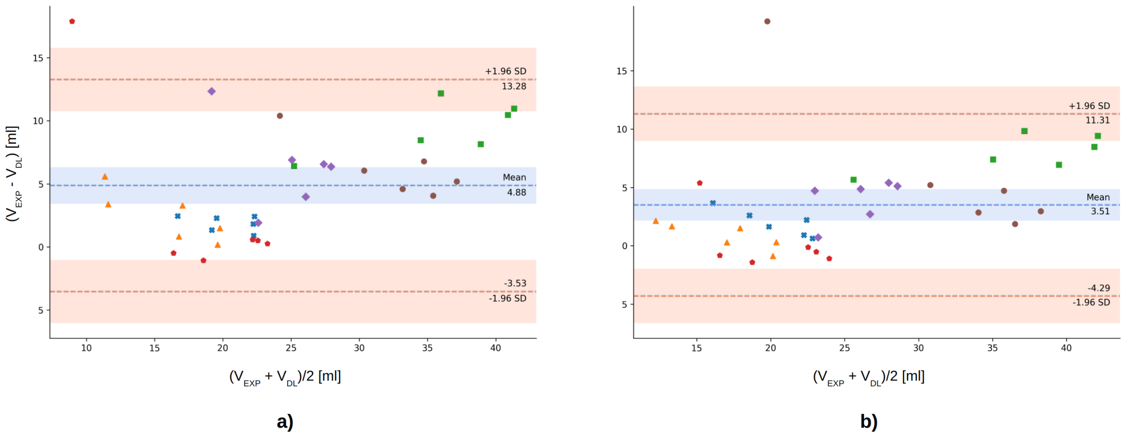


Fig. 2: Bland-Altman plots from Experiment 1 (**a**) and 2 (**b**) showing differences in measurements between MRI experts and deep learning (DL) ensembles on SUH data. Each color of the markers represents a unique study subject.

5. DATA AND METHOD AVAILABILITY

The source code is available at `https://github.com/MMIV-ML/fastMONAI/research`. We've shared the weights of our final neural network, enabling the application of our model to produce vertebrae masks in T2w MRI scans.

6. COMPLIANCE WITH ETHICAL STANDARDS

Written informed consent was obtained from all patients per the Helsinki Declaration. The study was approved by the Regional Ethics Committee in Norway (REC South East, reference number 2017/2450) and the Regionala Etikprövningsnämnden i Göteborg (reference numbers 888-14/483-17).

7. ACKNOWLEDGEMENTS

Trond Mohn Research Foundation (BFS2018TMT07) and the Norwegian national program for clinical therapy research, KLINBEFORSK (2017–201), supported our study.

8. REFERENCES

[1] NJ Sheehan, "Magnetic resonance imaging for low back pain: indications and limitations," *Annals of the rheumatic diseases*, vol. 69, no. 01, pp. 7–11, 2010.

[2] AS Lundervold and A Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[3] A Sekuboyina et al., "VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images," *Medical image analysis*, vol. 73, pp. 102166, 2021.

[4] Y Deng et al., "CTSpine1K: A Large-Scale Dataset for Spinal Vertebrae Segmentation in Computed Tomography," *arXiv preprint arXiv:2105.14711*, 2021.

[5] A Diaz-Pinto et al., "MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images," *arXiv e-prints*, 2022.

[6] JT Lu et al., "Deep Spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning," in *Machine Learning for Healthcare Conference*. PMLR, 2018, pp. 403–419.

[7] J Zhou et al., "Automatic Vertebral Body Segmentation Based on Deep Learning of Dixon Images for Bone Marrow Fat Fraction Quantification," *Frontiers in endocrinology*, vol. 11, pp. 612, 2020.

[8] KT Gao et al., "Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning," *JOR Spine*, p. e1204, 2022.

[9] Nikolas Lessmann et al., "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Medical image analysis*, vol. 53, pp. 142–155, 2019.

[10] Chengwen Chu et al., "Annotated T2-weighted MR images of the Lower Spine," July 2015.

[11] K Storheim, A Espeland, et al., "Antibiotic treatment In patients with chronic low back pain and Modic changes (the AIM study): study protocol for a randomised controlled trial," *Trials*, vol. 18, no. 1, pp. 1–11, 2017.

[12] PM Kristoffersen, A Espeland, et al., "Short tau inversion recovery MRI of Modic changes: a reliability study," *Acta radiologica open*, vol. 9, no. 1, pp. 2058460120902402, 2020.

[13] H Hebelka, H Brisby, K Lagerstrand, et al., "Axial loading during MRI induces significant T2 value changes in vertebral endplates—a feasibility study on patients with low back pain," *Journal of Orthopaedic Surgery and Research*, vol. 13, no. 1, pp. 1–7, 2018.

[14] Paul AY et al., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[15] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, p. 896.

[16] S Kaliyugarasan, M Kocinski, A Lundervold, and AS Lundervold, "2D and 3D U-Nets for skull stripping in a large and heterogeneous set of head MRI using fastai," in *NIK2020*, 2020.

[17] S Kaliyugarasan, A Lundervold, and AS Lundervold, "Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using fastai and MONAI," *IJIMAI*, 2021.

[18] E Hodneland, S Kaliyugarasan, et al., "Fully Automatic Whole-Volume Tumor Segmentation in Cervical Cancer," *Cancers*, vol. 14, no. 10, pp. 2372, 2022.

[19] J Howard and S Gugger, "Fastai: a layered API for deep learning," *Information*, vol. 11, no. 2, pp. 108, 2020.

[20] MONAI Consortium, "MONAI: Medical Open Network for AI," 2022.

[21] F Pérez-García, R Sparks, and S Ourselin, "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106236, 2021.

[22] O Ronneberger, P Fischer, and T Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] L Wright and N Demeure, "Ranger21: a synergistic deep learning optimizer," *arXiv preprint arXiv:2106.13731*, 2021.

[24] LN Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[25] N Abraham and NM Khan, "A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[26] S Sudirman et al., "Lumbar Spine MRI Dataset," 2019.