Addressing Data Heterogeneity in Image-Based Computer-Aided Detection and Diagnosis Methods

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der Technischen Wissenschaften

eingereicht von

Dipl.-Ing. Maria Wimmer
Matrikelnummer 0725248

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Zweitbetreuung: Dipl.-Math.in Dr.in Katja Bühler

Diese Dissertation haben begutachtet:

Timo Ropinski

Horst Karl Hahn

Wien, 22. Mai 2024

Maria Wimmer

TU WIEN Informatics

Addressing Data Heterogeneity in Image-Based Computer-Aided Detection and Diagnosis Methods

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Technischen Wissenschaften

by

Dipl.-Ing. Maria Wimmer
Registration Number 0725248

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Second advisor: Dipl.-Math.in Dr.in Katja Bühler

The dissertation has been reviewed by:

Timo Ropinski

Horst Karl Hahn

Vienna, 22nd May, 2024

Maria Wimmer

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Maria Wimmer

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 22. Mai 2024

Maria Wimmer

Acknowledgements

I am very grateful to my advisors Meister Eduard Gröller from TU Wien and Katja Bühler from the VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH. I thank Meister Eduard Gröller for supervising this thesis and for his valuable advice and comments. Thanks to Katja Bühler for giving me the opportunity to conduct research in the exciting field of medical image analysis at VRVis and for her guidance and constructive feedback throughout this journey. And I want to express my gratitude to Timo Ropinski and Horst Hahn for reviewing this thesis.

This work would not have been possible without all the great current and former colleagues and friends in the Biomedical Image Informatics group and at VRVis as a whole. Special thanks to my co-authors David Major, Alexey Novikov, Dimitrios Lenis, Gert Sluiter, Astrid Berg, and Theresa Neubauer, for fruitful discussions and constructive feedback, and your support in stressful deadline situations. Thanks to my initial project colleagues Alexey Novikov, David Major, and Jiří Hladůvka, for the many discussion sessions on and off topic and for insightful and fun conference trips. In particular, I owe a big thank you to David Major, who has accompanied me since my very first day at VRVis. Thank you for countless brainstorming sessions and for reviewing my publications and this thesis. Moreover, thanks for always having an open ear and thoughtful advice when needed. I also like to thank Astrid Berg and Theresa Neubauer for all the wisdom and laughter we share in our office. Thanks for keeping the chocolate supply at a good level and for sending cute cat and alpaca pictures. And I want to thank all the other great colleagues and friends I found at VRVis, for making not only lunch and coffee breaks much more enjoyable but also activities outside the office. Thank you for your thoughtful words and emotional support whenever I needed advice and someone to talk to. This belongs to many, but in particular to Michi Schwärzler, Harri Steinlechner, Hendrik Schulze, Thomas Ortner, Didi Drobna, Irmi Fuchs, and Andi Buttinger-Kreuzhuber. Special thanks to Irmi Fuchs for proofreading large parts of this work.

I also like to thank my friends in my hometown and all the friends I found in Vienna in the student dorm, at the university, and VRVis. Thanks for not only distracting me by going out for a drink, visiting concerts and festivals, going for ice cream, wine tastings, or numerous hiking trips, but also for cheering me up and bringing me back on track whenever I was unsure whether to pursue my doctoral studies.

A special thank you deserves the MIST – the Medizinischer Informatik Stammtisch – for our fun regular get-togethers and weekend trips, which we should do more often again. I am especially grateful to Astrid, Marie-Luise, and Sabrina for being a significant part of my life since my childhood. Thank you for all the good times we share and for always being there for me whenever I need help. I know that I can always count on you. And I am deeply thankful to Edith for being by my side for more than half my life. Thank you for your tremendous support and thoughtful words in every life situation. I am very fortunate to have you as a friend.

I am very grateful to all my family, especially my parents, Maria and Franz, who made it possible for me to study in the first place, and my brother Christian. Thank you for believing in me and always encouraging me on my way, even though you often wondered what kept me from visiting more often.

Last but not least, I want to thank Martin with all my heart. Thank you for your endless support and for providing me with food and chocolate in desperate times of paper and thesis writing. Most of all, thank you for your patience, understanding, and emotional support, especially in the final stages of this thesis. I could not have finished this project without you!

This work was enabled by the Competence Center VRVis. VRVis is funded by BMK, BMDW, Styria, SFG, Tyrol, and Vienna Business Agency in the scope of COMET – Competence Centers for Excellent Technologies (879730) which is managed by FFG.

Kurzfassung

Die Aufnahme von medizinischen Bilddaten ist ein notwendiger Schritt im Rahmen von Vorsorgeuntersuchungen, für die Diagnose von Krankheiten, Planung von Operationen oder Therapien sowie für die Überwachung von Krankheitsverläufen. In der klinischen Praxis werden die Daten von medizinischen Experten analysiert und befundet. Dies kann eine sehr zeitaufwändige Aufgabe sein. Daher wird seit Jahrzehnten intensiv an der automatisierten Analyse medizinischer Bilddaten geforscht und an der Frage, wie die zuvor genannten Aufgaben durch computergestützte Detektions- und Diagnosealgorithmen unterstützt werden können. Eine der größten Herausforderung in diesem Zusammenhang ist die hohe Heterogenität medizinischer Bilddaten. Die Aufnahme von Daten mit verschiedenen bildgebenden Verfahren, wie z.B. Röntgen oder Magnetresonanztomographie (MRT), die Änderung von Aufnahmeparametern und der Einsatz verschiedener Scanner führen zu einem diversen Set an Daten. Die unterschiedliche räumliche Auflösung sowie die hohe Dimensionalität der Daten stellen zusätzliche Herausforderungen bei der Entwicklung automatisierter Lösungen dar.

In dieser Dissertation untersuchen wir Machine Learning-basierte Methoden für die Analyse heterogener medizinischer Bilddaten, wie z.B. multi-parametrische Daten, multimodale Daten, Daten aus unterschiedlichen Aufnahmerichtungen oder aus verschiedenen Krankenhäusern. Wir stellen drei verschiedene Analyse-Pipelines vor, die *generalisierende* und auf *Fusion* basierende Ansätze anwenden, und demonstrieren ihre Anwendbarkeit auf unterschiedlichen öffentlichen Datensätzen. Unsere Methoden adressieren zwei ausgewählte Anwendungsfälle in der Radiologie: die *semantische Annotation der Wirbelsäule in MRT Daten* und die *Analyse von Mammografie-Bildern*.

Ein Problem bei der semi- und vollautomatischen Annotation der Wirbelsäule in MRT-Daten ist die Tatsache, dass MRT-Bilder keine standardisierte Intensitätsskala aufweisen. Dies führt zu einer großen Vielfalt an unterschiedlichen Bildkontrasten. Wir schlagen deshalb eine iterative Lösung vor, die Entropy-Optimized Texture Models (ETMs) verwendet. Die Anwendung von ETMs ermöglicht uns, die trainierten Modelle auf eine große Bandbreite unterschiedlicher MRT-Daten anzuwenden. Dieser Ansatz steht im Gegensatz zu verschiedenen Lösungen aus der Literatur, die Methoden für spezifische MRT-Sequenzen und -Protokolle entwickeln.

Im Rahmen von Mammografie-Vorsorgeuntersuchungen werden pro Patientin bzw. Patient nicht nur einzelne Röntgenbilder aufgenommen, sondern vier Bilder aus unterschiedlichen

Aufnahmerichtungen. Diese werden zu einer Studie zusammengefasst, die für eine bildgestützte Analyse zur Verfügung stehen. Zusätzlich zu diesen Bildern sind Informationen auf verschiedenen Ebenen vorhanden, z.B. auf Patienten-, Bild- oder Läsionsebene. Um diese Informationen effizient zu nutzen und zu kombinieren, entwickeln wir mehrere Deep Learning-basierte Modelle. Diese behandeln bestimmte Aufgaben, die bei der Befundung von Mammogrammen wichtig sind, wie z.B. die Lokalisierung von Abnormalitäten. Für eine verbesserte Vorhersage auf Patientenebene fusionieren wir Ergebnisse und extrahierte Merkmale der einzelnen Modelle, um die Performanz zu erhöhen. Dieser Ansatz steht im Gegensatz zu Methoden, die einfache Kombinationsansätze verwenden.

Die in dieser Dissertation präsentierten Ergebnisse zeigen, dass die Berücksichtigung der verschiedenen Aspekte heterogener medizinischer Bilddaten unumgänglich ist, um sowohl die Generalisierungs- als auch die Vorhersagefähigkeit computergestützter Detektions- und Diagnosemethoden zu verbessern.

Abstract

The acquisition of medical imaging data is inevitable for screening, diagnosis, planning of surgery or therapy, or monitoring of diseases. In clinical practice, the data is assessed by medical experts, which can be a very time-consuming task. Hence, for decades a lot of research effort has been dedicated to the automated analysis of medical imaging data and to the question of how Computer-Aided Detection and Diagnosis algorithms can assist the tasks mentioned above. However, one of the biggest challenges in this regard is the highly heterogeneous nature of medical imaging data. The acquisition of data from different imaging modalities, like X-ray or Magnetic Resonance Imaging (MRI), changes of acquisition parameters, and the use of different scanners results in diverse data. The varying spatial resolution as well as the high dimensionality of the data pose additional challenges to the development of automated solutions.

In this thesis, we investigate different machine learning-based methods to address the analysis of heterogeneous medical imaging data, such as multi-parametric, multi-modal, multi-center, or multi-view data. We present three different pipeline approaches that follow *generalization*- and *fusion*-based approaches and demonstrate their applicability on diverse public datasets. Our contributions target two selected use cases in radiology: the *semantic labeling of the spine in MRI data* and the *analysis of mammograms*.

In semi- and fully-automated spine labeling in MRI data, we are confronted with the problem that MRI data does not exhibit a standardized intensity scale, which results in a large variety of different image contrasts. To overcome this problem for semantic spine labeling, we propose an iterative labeling pipeline that employs Entropy-Optimized Texture Models (ETMs). The application of trained ETMs allows us to apply our models to a wide range of different MRI data. This is in contrast to various related works that develop methods for specific MRI image sequences and protocols.

For the analysis of mammography screening data, not only one but four X-ray images from different fields of view are available that form a study of a patient. In addition to this multi-view data, we deal with multi-scale information at various levels, e.g., on a patient, image, or lesion level. To utilize and combine this information efficiently, we develop several deep learning-based models that aim for a specific task important in examining mammograms, such as the localization of abnormalities. For a comprehensive prediction on a patient level, we propose to fuse predictions and features from the individual models to increase performance, which is in contrast to standard ensembling techniques.

The results in this thesis demonstrate that considering the different aspects of heterogeneous medical imaging data is inevitable to improve both generalization and predictive capabilities of Computer-Aided Detection and Diagnosis methods.

Contents

“If it weren’t for the last minute, nothing would get done.”

Rita Mae Brown, Writer and Activist

CHAPTER 1

Introduction

1.1 Clinical Motivation

In every person's life, there may come a time when there is a need to acquire medical imaging data to obtain insights into a specific medical condition. Depending on the indication and clinical question, different types of imaging data will be acquired, e.g., X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), or Positron Emission Tomography (PET) scans. In clinical practice, image acquisition is part of the *diagnostic imaging workflow*, which comprises different stages. It includes all steps ranging from the initial image ordering from the referring physician to the scheduling of the exam and image acquisition itself, continues with the image reading and interpretation by the radiologists, and is concluded by reporting and communication of results to the referring physician and patient [105, 180]. Figure 1.1 illustrates the steps.

Over the last decade, the number of referrals to a medical image acquisition procedure and, thus, the amount of acquired imaging data increased steadily, as shown in Table 1.1. This development has driven the effort to increase the efficiency along the whole workflow, especially through machine learning-based approaches. Numerous methods have been presented that optimize the scheduling of exams, accelerate image acquisition, increase

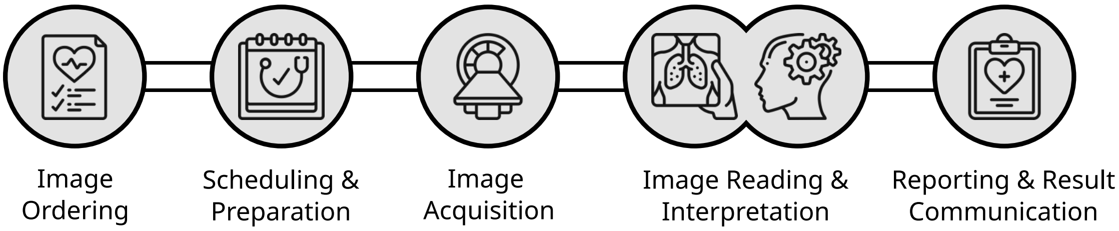


Image Ordering
Scheduling & Preparation
Image Acquisition
Image Reading & Interpretation
Reporting & Result Communication

Figure 1.1: Main stages of the diagnostic imaging workflow, adapted from Preim & Botha [180]. Icons made by Smashicons from www.flaticon.com.

		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
CT	AUT						175.2	172.6	173.4	183.6	196.2	182.0
	GER	120.6	127.5	131.2	135.3	143.8	143.1*	148.5*	139.9*	144.7*	151.2*	150.0*
	US	264.8	273.8	256.8	240.5	255.0	245.4	253.8	256.2	271.4	278.4	220.2
MRI	AUT						117.4	120.2	130.7	141.4	148.0	140.5
	GER	105.5	110.5	115.3	124.2	131.3	138.6*	143.4*	140.8*	145.1*	149.8*	149.9*
	US	97.6	102.7	104.8	106.9	109.6	117.9	120.7	111.0	119.3	127.9	82.7
PET	AUT						2.1	2.4	2.4	4.8	4.9	4.6
	GER	1.4	1.6	1.6	1.6	1.6	1.6	1.7	1.7	1.8	1.9	1.8
	US	5.6	5.9	5.5	5.0	5.1	5.4	5.8	6.0	6.4	6.7	6.7

Table 1.1: Number of CT, MRI, and PET exams per 1.000 population in Austria (AUT), Germany (GER) and United States (US) from 2010 to 2020. Numbers marked with * indicate estimates. Data from OECD [166].

image quality, or create reports automatically [105, 189]. One of the largest fields of research along the workflow is the automated analysis of medical imaging data, like radiology or oncology data [84, 183]. The development of (semi-)automated Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) methods in radiology already have a long history dating back to the 1950s and early 1960s [55]. Due to the increasing amount of imaging data and, thus, increasing workload for radiologists, this development continues unabated.

1.2 Problem Statement

1.2.1 Heterogeneous Medical Imaging Data

Not only the number of acquired scans is rapidly growing, but also the complexity of medical data is increasing, which is one of the main challenges in medical image analysis [148]. One influencing factor is, for example, the advances in medical imaging techniques that result in higher-resolution data [182]. In the literature, the complexity of medical (imaging) data is further explained by the *heterogeneous nature* of the data [109, 148, 182]. Kehrer & Hauser [109] and Raidou [182] describe the different aspects of heterogeneous scientific data in general and medical data specifically as *multi-faceted*. Thereby, they refer to the heterogeneity characteristics of the data, e.g., the *multi-modal* or *spatio-temporal* nature.

For medical imaging data, we can identify various heterogeneity characteristics of the data, which we will refer to as *categories of heterogeneous data* in this thesis. We can group them into the following *categories* $\mathbf{c}^i, i = 1, \ldots, m$, where each comprises various *types* $\mathbf{t}^j, j = 1, \ldots, n$:

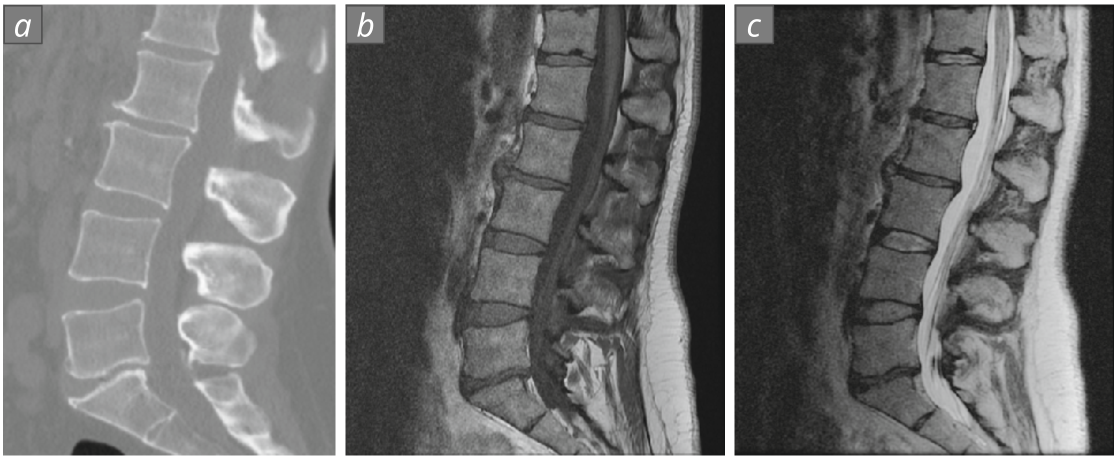


Figure 1.2: Sample images from the same patient showing the spinal column acquired with two different imaging modalities (*multi-modal*) and varying scanning parameters (*multi-parametric*): (a) CT scan, (b) T1w FLAIR MRI scan, (c) T2w CUBE MRI scan. Images provided by Cai et al. [28], downloaded from Spineweb [215].

Multi-modal data. Multi-modal data refers to data from various acquisition modalities, i.e., types $\mathbf{t}^j$, such as CT, MRI, or X-ray [109]. One essential property of multi-modal data is their *complementarity*, as described by Lehat et al. [117]. They state that "*each modality brings to the whole some type of added value that cannot be deduced or obtained from any of the other modalities in the setup*".

Multi-parametric data. Changes in the image acquisition parameters, e.g., in MRI, result in medical images of the same modality that exhibit different image contrasts. In MRI, this is also often referred to as *multi-sequence data*. Different types are for example T1-weighted (T1w) or T2-weighted (T2w) data. Figure 1.2 shows multi-modal, multi-parametric sample images of the spinal column from the same patient.

Multi-dimensional data. The dimensionality of the data usually depends on the acquisition modality, e.g., 2D X-ray data or 3D data from MRI, CT, or PET scanners [148, 179].

Multi-resolution data. Depending on the region of interest and the image acquisition process and devices, we are concerned with limited and varying resolution of the data. A related issue are highly anisotropic voxel sizes, i.e., where the pixel size within a slice can be several times smaller than the slice distance [179].

Multi-scale data. By multi-scale data, we refer to the different levels of the data, e.g., at the molecular level or cellular/tissue level [177]. On an image level, we also refer to multi-scale data if we speak, for example, about a Region of Interest (ROI) within an image vs. the full image or data on a patient level. Further, also annotations can occur at multiple scales.

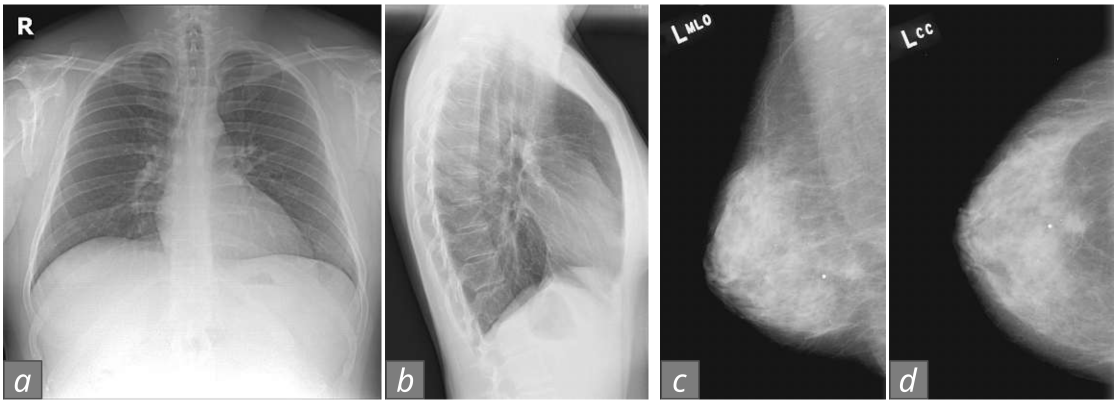


Figure 1.3: Sample X-ray images showing the chest and breast from different types of views: (a) anterior-posterior and (b) lateral chest X-ray, (c) Mediolateral oblique (MLO) and (d) Craniocaudal (CC) mammogram. Chest images from PadChest dataset [24], Mammography images from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset [121, 122].

Multi-view data. By multi-view data, we refer to data that comprises different fields of view of the same anatomical structure or region of interest in one patient. The acquisition of multi-view data is for example common clinical practice in breast imaging for X-ray mammography and 3D digital breast tomosynthesis. In addition, 3D imaging modalities allow physicians to inspect the same region from different viewpoints intrinsically. With MRI, data can even be acquired in arbitrary fields of view. Figure 1.3 shows sample multi-view X-ray images of the chest and breast.

Multi-temporal data. This category refers to data from one subject at multiple time steps, e.g., in follow-up studies. Another aspect of temporal data is *spatio-temporal* data where for given spatial locations temporal data is recorded, e.g., electroencephalography or functional MRI data in spatio-temporal brain imaging [107].

Multi-subject data. Multi-subject data refers to data that has been acquired from different individuals.

Multi-vendor and Multi-center data. Finally, multi-vendor data refers to imaging data acquired with scanners from different vendors, e.g., Siemens, Philips, while multi-center data relates to data from different hospitals or imaging facilities.

Aside from the different categories of data heterogeneity, one factor that influences the heterogeneity of the data and the regions and objects of interest in the data, is the acquisition process itself – and its limitations. Examples are limited data resolution, noise, or the presence of imaging artifacts. They can originate from the image reconstruction algorithms, or from foreign objects such as implants, from low X-ray doses or due to motion artifacts caused by patient movement [68, 148]. Finally, not all imaging modalities are standardized. While CT data is normalized to Hounsfield Units, MRI data does not follow a standardized intensity scale. Hence, the intensity ranges of the same tissues vary within a scan and across scans [179]. These issues are further amplified if different medical scanners are used or the data is acquired in different hospitals, i.e., in *multi-center* data scenarios.

1.2.2 Medical Problem Statement

In clinical practice, a radiologist views, analyzes, and interprets heterogeneous imaging data as part of the diagnostic imaging workflow, as illustrated in Section 1.1 and Figure 1.1. The outcome of this process is a *single* result, that is, a medical diagnosis. The amount of image data the radiologist must examine when reading a patient varies. In some cases, only a *single* 2D or 3D image is available for analysis. However, more commonly, the doctor has to examine a *set of image data* with different heterogeneity characteristics. For example, the set may include data from different views and modalities or one or multiple prior studies. There are two different ways how the data can be treated: Radiologists can look at the data (a) separately or (b) fuse it prior to the image reading. In the first case, the data is viewed separately or side-by-side. Common radiology software enables the doctor to view different images simultaneously side-by-side or register the data and link the different views accordingly. In this case, the radiologist fuses the data and/or result "mentally" to obtain a single result. In the second case, the heterogeneous data is fused with dedicated methods [80]. For example, anatomical and functional imaging modalities like MRI and PET, respectively, may be combined, or prior and current studies. The fused data is then examined by the radiologist to derive the single result.

1.3 Scope of the Thesis

In this thesis, we target the heterogeneous nature of medical imaging data from a technical point of view in the development of CADe and CADx methods and propose different ways to approach it. While some categories, like, for example, multi-subject data, are usually not explicitly addressed, others, like the multi-dimensionality aspect, require more attention as they bring various challenges with them.

Given we want to develop a learning-based method $\mathbf{M}$ for a heterogeneous medical image dataset $\mathcal{D}$ that comprises data of different types $\mathbf{t}^j$, e.g., a multi-modal dataset with CT and MRI data. We denote the subsets of $\mathcal{D}$ with a given type $\mathbf{t}^j, j = 1, \ldots, n$ and category $\mathbf{c}^i, i = 1, \ldots, m$ as $\mathbf{D}_i^j$, with $\mathbf{D}_i^j \in \mathcal{D}$. We can identify three general approaches how to treat the different subsets $\mathbf{D}_i^j$ in the development of method $\mathbf{M}$ to obtain a result $\mathbf{R}$. They are illustrated in Figure 1.4 and listed in the following:

(a) **Standard approach:** We can develop n methods $\mathbf{M}^j$ for each type of data $\mathbf{D}_i^j$, $j = 1, \ldots, n$ that are only applicable to the specific type $\mathbf{t}^j$ where they have been trained on and predict n type-specific results $\mathbf{R}^j$.

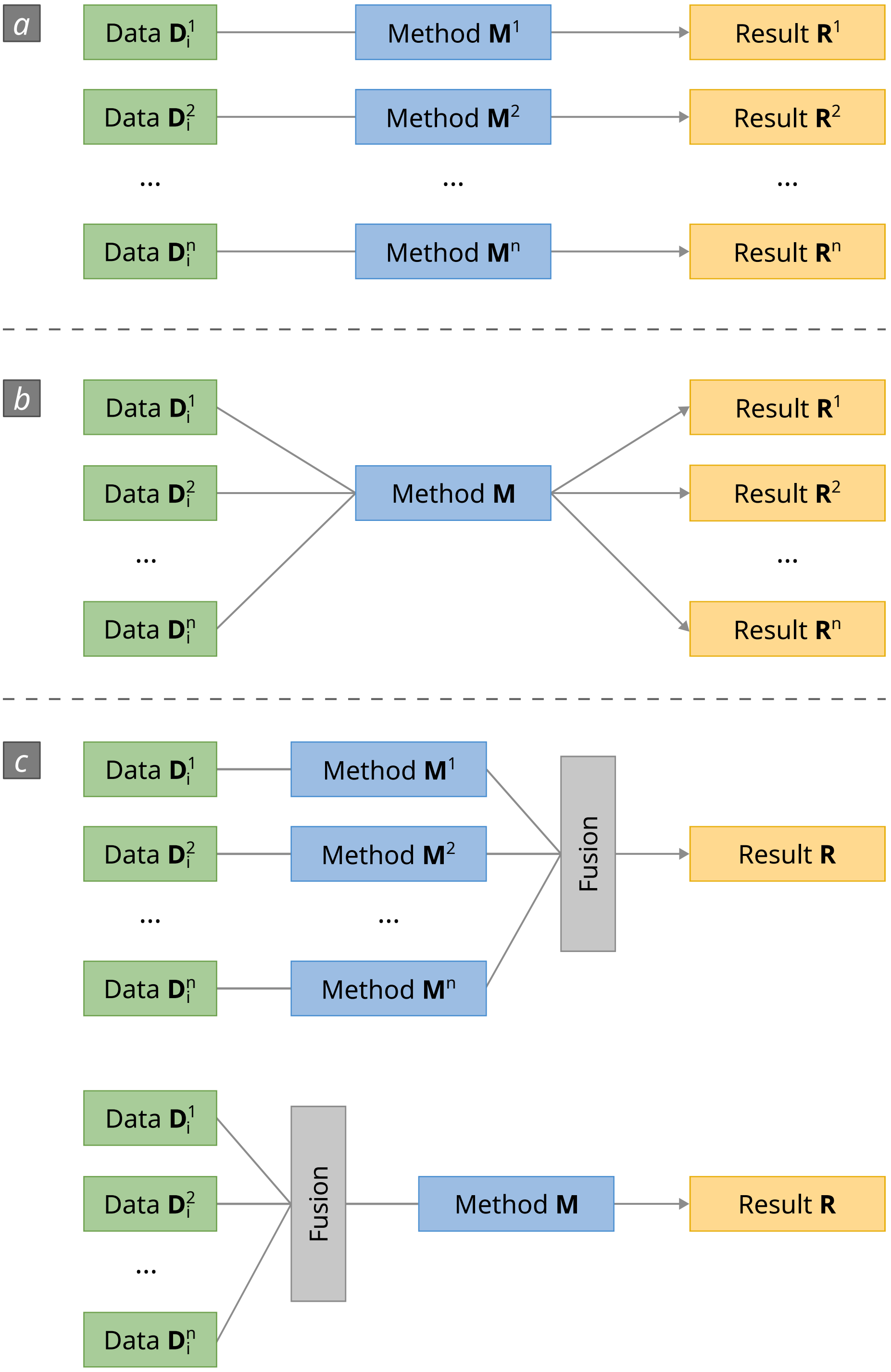


Figure 1.4: Three general approaches to deal with a heterogeneous dataset comprising different types $\mathbf{t}^j$: (a) Standard approach, (b) Generalization approach, (c) Fusion approach.

(b) **Generalization approach:** *One* method $\mathbf{M}$ is developed that operates independently / separately on data $\mathbf{D}_i^j, j = 1, \ldots, n$ and predicts one result $\mathbf{R}^j$ per type $\mathbf{t}^j$ separately. This enables independent processing of the data even if not all types $\mathbf{t}^j$ are available.

(c) **Fusion approach:** For this approach, we derive two general settings: (1) we can have multiple methods $\mathbf{M}^j$ for each type $\mathbf{t}^j$ that operate on the specific type of data $\mathbf{D}_i^j, j = 1, \ldots, n$ where they have been trained on but in the end predict only *one* combined result $\mathbf{R}$. Depending on where the data is combined, we refer to this setting as *intermediate* or *late* fusion. (2) *One* method $\mathbf{M}$ that processes the heterogeneous data $\mathbf{D}_i^j$ *simultaneously* and predicts one general result $\mathbf{R}$. In this case, *early* fusion is performed.

We note that combinations of these general approaches are possible and very common in practice, especially for (b) and (c). One example is the development of fusion techniques for different types of multi-modal data, while generalization-based approaches alleviate the need to deal, e.g., explicitly with multi-vendor characteristics.

In this thesis, we specifically focus on (b) and (c) and develop fusion- and generalization-based methods that target the following categories of heterogeneous medical imaging data: multi-modal, multi-parametric, multi-vendor, multi-subject, multi-center, multi-scale, and multi-view data.

1.4 Related Work

In the following, we give a brief overview of the main concepts of how data heterogeneity can be addressed in the context of medical image analysis. We will focus on learning-based methods for CADe and CADx and provide different examples from the literature.

In general, we can split the methods into those based on a specific model, i.e., *model-based approaches*, and those applicable independently of the underlying model, i.e., *model-agnostic approaches*. While both types will be part of the following review, we structure it as follows: Section 1.4.1 discusses methods following the generalization-based approach in category (b) that deal with heterogeneous data individually with a single method. In Section 1.4.2, we review fusion-based concepts that fall in category (c) and discuss how we can treat the data jointly with single or multiple methods. We note that we will not discuss specific approaches for variant (a) in this section. It is the most common variant to develop type-specific approaches for a specific medical question. We refer to Section 2.1.1, Section 3.1.1, and Section 4.1.1 for an in-depth review of related works that focus on specific categories of heterogeneous data for the two selected CADe and CADx application parts of this thesis.

1.4.1 Generalization-Based Approach

Approaches that aim to tackle the data heterogeneity with a single method **M** are usually concerned with increasing the *generalization* capability of the method. In medical image analysis, this is of high interest, especially for the multi-modal, multi-parametric, multi-sequence, multi-vendor, multi-center categories of heterogeneous data.

One way to achieve generalization capabilities is through data preprocessing strategies that normalize or standardize intensity values before the actual CADe or CADx task. Due to the heterogeneous nature of MRI data, a lot of methods have been proposed for this purpose [190, 210, 244]. Another approach presented before the current era of deep learning by Zambal et al. [260] combines data normalization and segmentation of anatomical structures in 2D data in a single model. They propose a model-based approach called Entropy-Optimized Texture Models (ETMs) that extend Active Appearance Models (AAMs) with a texture normalization method based on entropy. It allows building models that are robust against texture variations, e.g., brightness or low contrast, and even different imaging modalities. This enables the usage of ETMs for multi-parametric or multi-modal data.

In the current deep learning era, a lot of research effort is put into methods independent of a specific model, i.e., *model-agnostic* approaches. On the one hand, the learning and generalization capabilities of deep learning models have increased, due to sophisticated learning strategies, data preprocessing and augmentation, loss functions, and regularization techniques. These developments and strategies – to name a few – already alleviate to a certain extent the need to treat some of the heterogeneity aspects explicitly during model development. Approaches such as the No New-Net [92, 93] show that elaborate data preprocessing, like standardizing MRI data, as well as extensive data augmentation, and choice of training regime, can be more effective in training a standard U-Net compared to complex architectural modifications. Another factor is data diversity, as it is the case with clinical routine data, which is considered more important than the choice of the underlying deep learning model in a recent study [82].

On the other hand, research fields dedicated to increasing the generalization capability have gained momentum in the medical domain, for example, *continuous learning* or *domain adaptation* [72, 178]. In both fields – among others – we are concerned with the problem of *domain shifts* between one or multiple domains. The term “domain shift” refers to changes in the data that result in a shift from the initial data distribution, e.g., due to changes in the image acquisition or data from another modality.

Continuous Learning. In *continuous learning*, we train a model sequentially on data from new domains, i.e., where the data distribution changes or new tasks or classes become available over time [230]. While in domain adaptation, the main goal is to maximize performance on the target domain(s), in the field of continuous learning, the performance on *all* previously seen data distributions or tasks should be maintained. This is in contrast to naive sequential learning, which suffers from the problem of *catastrophic*

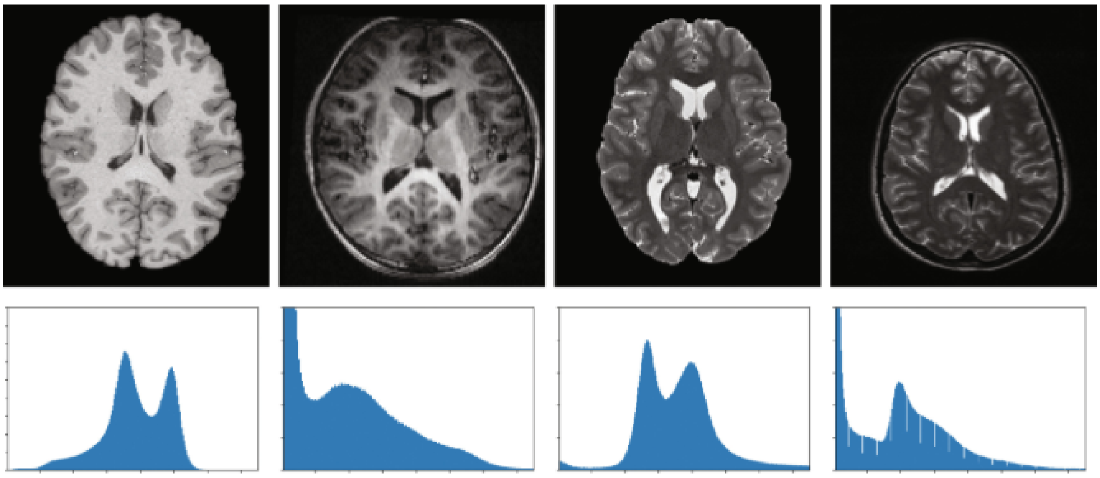


Figure 1.5: Axial T1w and T2w MRI images acquired with different scanners (top row) and corresponding histograms (bottom row). The first and third image (top row) are skull-stripped MRI images. Image from Karani et al. [106].

inference or *catastrophic forgetting*, that is, the performance degradation of a model on “old” data and tasks [153].

Li et al. [130] propose a so-called style-oriented replay module to improve heart segmentation in MRI data. The module generates imaging data from the base data and all previously seen data domains instead of storing the original data. During continuous training, style parameters corresponding to the new dataset are learned in addition to the optimization of their multi-class segmentation model. Additionally, they perform feature whitening to reduce the sensitivity to changes in the data domain and to increase the generalization capability to unseen data. The method by Karani et al. [106] learns variations in multi-center, multi-vendor, multi-parametric MRI brain data with domain-specific parameters in the batch-norm layers of their model for brain segmentation. A “domain” refers to a certain Magnetic Resonance (MR) protocol in this context. Figure 1.5 shows a sample of their data. In the continuous training steps, only the domain-specific batch normalization layers are updated while the learned convolutional filters stay fixed. Unlike the work by Karani et al. [106], other methods do not require knowledge about the current domain in a continuous training setting [81, 130, 176, 218]. This is much closer to a real-world clinical setting. Another important factor in clinical applications is data privacy, which is why methods that do not store any kind of data are preferred [106, 130] over those that store full images or – as a trade-off – feature representations in their memory [81, 176, 218]. Hofmanninger et al. [81] propose a method that dynamically maintains a sample memory. The memory is updated after each training step by calculating the stylistic differences of new and already-seen images based on the Gram matrix. Their method can deal with a continuous shift between different CT acquisition protocols and also in the appearance of the classification target related to the class of interest. In a more recent work, they extend their method with a pseudo-domain detection module that further increases the diversity of the memory [176]. Srivastava et al. [218] utilize a replay-based method that replays compressed, intermediate feature representations instead of full images. They evaluate their method for a domain shift scenario in multi-label chest X-ray pathology classification between three public datasets.

Domain Adaptation and Generalization. In *domain adaptation* – a special variant of transfer learning – a model trained for a fixed task, e.g., organ segmentation, on a dataset from a source domain is applied at test time on a set of data from another domain – the target domain. One example is the application of the model on data from a different medical scanner or another modality [72]. This is the simplest scenario in domain adaptation, whereby various other directions, such as multiple adaption steps, or generalization from various source domains to one or many target domains are also investigated. The most general variant is *domain generalization* where neither labeled nor unlabeled data from the target domain is available. We review a selection of approaches in the following and refer to the literature for a comprehensive overview [72].

The lack of annotated training data is a common issue in medical image analysis that fosters the development of unsupervised domain adaptation methods, i.e., where the data of the target domain is unlabeled. Various methods have been proposed, for example, for unsupervised segmentation tasks in cross-modality or cross-center/cross-vendor settings [58, 104, 128, 209, 264]. Kamnitsas et al. [104] present a 3D approach for brain lesion segmentation in multi-sequence MRI scans. They jointly train the segmentation network as well as a domain discriminator that classifies the domain of the current training data. In their adversarial training strategy, they aim to align the feature spaces of the source and target domain. Dou et al. [58] follow a similar approach for multi-modal data. They train a multi-class segmentation network on MR data of the heart as a first step. In the subsequent adversarial training, the lower-level features of the segmentation network adapt to the target domain comprising CT scans, while the higher-level features are frozen and shared between both domains. Their proposed model comprises two discriminators, one that discriminates features from the source and target domain, while the second one constrains the segmentation masks to fulfill shape requirements. While the previous works employ a feature alignment strategy, Shin et al. [209] investigate a domain translation-based approach for adaptation from T1-weighted (T1w) to T2-weighted (T2w) MRI data. First, they generate pseudo-T2w images, constrained by the segmented structures in the T1w source domain image data. Then, in an iterative training, they employ the pseudo T2w data and real labels, segment the real T2w data, infer pseudo labels and iterate training with both, real and pseudo T2w data and labels. For the actual segmentation, they adopt the nnU-Net approach [92].

One drawback of the presented domain adaptation methods is that source and target domain data must be available for the adversarial training strategies [58, 104, 209]. This is counteracted in the work of Zhang et al. [264], who investigate a domain *generalization* approach for 3D segmentation. Inspired by the common use of data augmentations, they apply a sequence of stacked image transformations during training. They adapt the quality, appearance, as well as spatial configuration of the training data that stems only from a single source domain. The authors validate their method on data that exhibits multi-vendor, multi-parametric, and multi-center characteristics whereby an adaptation between modalities is not tested. Inspired by the work on stacked image transformations, Li et al. [128] argue that linear dependencies between features from multiple source domains exist. They try to model these dependencies in the feature space while also learning a shared representation, i.e., aligning the distributions from the different source domains. The authors validate their domain generalization approach for classification and segmentation of multi-center, multi-vendor data for skin lesion classification as well as gray matter segmentation in the spinal cord.

Other Related Approaches. We also mention the related field of *disentangled representation* learning, which aims to *"separate out the main factors of variation that are present in our data distribution"*, as described by Liu et al. [138]. Chartsias et al. [33, 34] present several works in this field that aim to decouple spatial information from style information, i.e., anatomical from modality-related information. They demonstrate their semi-supervised approaches for different multi-class segmentation tasks in multi-modal and multi-parametric data. Overall, their method follows an autoencoder design, with several networks for the different tasks [33], i.e., separate encoders for modality and anatomy information, a decoder that reconstructs the input, and a decoder that performs the actual segmentation task.

1.4.2 Fusion-Based Approach

When it comes to the simultaneous/combined handling of heterogeneous data with *single* or *multiple methods*, we are concerned with methods and approaches that perform some form of *fusion* of heterogeneous image information. Baltrusaitis et al. [20] propose to classify multi-modal fusion approaches in general into *model-agnostic* and *model-based* approaches. By model-based approaches, they refer to kernel-based methods, graphical models, and neural networks. The model-agnostic fusion approaches are commonly coarsely divided into the *level* where the fusion happens: *early*, *intermediate*, and *late fusion*. These levels are also often referred to as *input-level*, *layer-level*, and *decision-level* fusion in the context of deep learning-based fusion strategies [270]. We will use the following terminology in this thesis, based on Stahlschmidt et al. [219] and Zhou et al. [270] (see Figure 1.6):

- **Early, input-, or image-level fusion:** In this variant, the raw input data $\mathbf{D}_i^j$ of the different types $\mathbf{t}^j$ is concatenated, and no representations of the data are learned before.
- **Intermediate, layer-, or feature-level fusion:** For each type $\mathbf{t}^j$ of the data, representations are learned before the fusion. In the case of neural networks, intermediate fusion can occur at one layer or gradually at different layers.

- **Late or decision-level fusion:** This is defined as the fusion or combination of decisions that are obtained by sub-methods.

There is also the possibility to mix the different variants and fuse at different levels. We will refer to this as *hybrid fusion* [20]. In general, there is no universal recipe on how and where the fusion should be performed to achieve the best results. Thus, oftentimes various combinations and fusion levels are explored in the literature to boost the performance. While early fusion is commonly used for its simplicity, the relationship between different modalities – in the case of multi-modal data – can be utilized only in a limited way. Intermediate fusion, on the other hand, exploits the connections between the learned representations from different branches or models, which usually benefits model performance. Further, it is very flexible in terms of where the fusion happens. Decision-level fusion usually leads to good feature representations per type of data but at the cost that connections between the different types cannot be utilized. Moreover, training different models is computationally expensive [219, 270].

The fusion of heterogeneous medical imaging data is commonly applied where multi-modal, multi-view, or multi-parametric data is present. Popular examples are organ segmentation [57, 108], tumor segmentation [162, 269], or mammography [101]. Especially for multi-modal and multi-parametric data, the *complementarity* property of this kind of data can be exploited efficiently with fusion approaches, according to Lahat et al. [117]. In the context of multi-view learning, there are also approaches that leverage Implicit Neural Representations (INRs) [155, 173, 213], as for example the popular Neural Radiance Fields (NeRFs) [157]. INRs and NeRFs found a wide adoption also in the medical domain, e.g., for synthesis of missing information, super-resolution tasks, reconstruction, registration, or novel view synthesis [159].

We note that within this thesis, we limit the discussion to approaches that perform the fusion with the primary goal of increasing the model performance of a higher-level task, e.g., segmentation or classification of certain anatomical structures or medical conditions. We refer to the literature for methods that focus solely on generating a fused image/volume from one or multiple modalities, e.g., to enhance or enrich image quality and content for subsequent image interpretation [17, 80, 97, 265].

Early and Intermediate Fusion. Dolz et al. [57] propose the HyperDense-Net for multi-sequence brain tissue segmentation, inspired by DenseNet [87]. Their network encodes features separately for each MRI sequence, namely T1w MRI and T2w MRI. Figure 1.7 illustrates a part of their architecture. Each image type has its own encoding path. The dense connections within each path and multiple connections across the two paths ensure multi-sequence feature learning and reduce the risk of overfitting. Hence, the fusion is performed at several stages in intermediate layers. They also extend this method for the segmentation of intervertebral discs in Dixon MRI data [56]. In the Dixon MRI sequence [62], four different image channels are acquired in the course of one image acquisition step [62]. Das et al. [50] and Li et al. [132] also segment discs in Dixon data

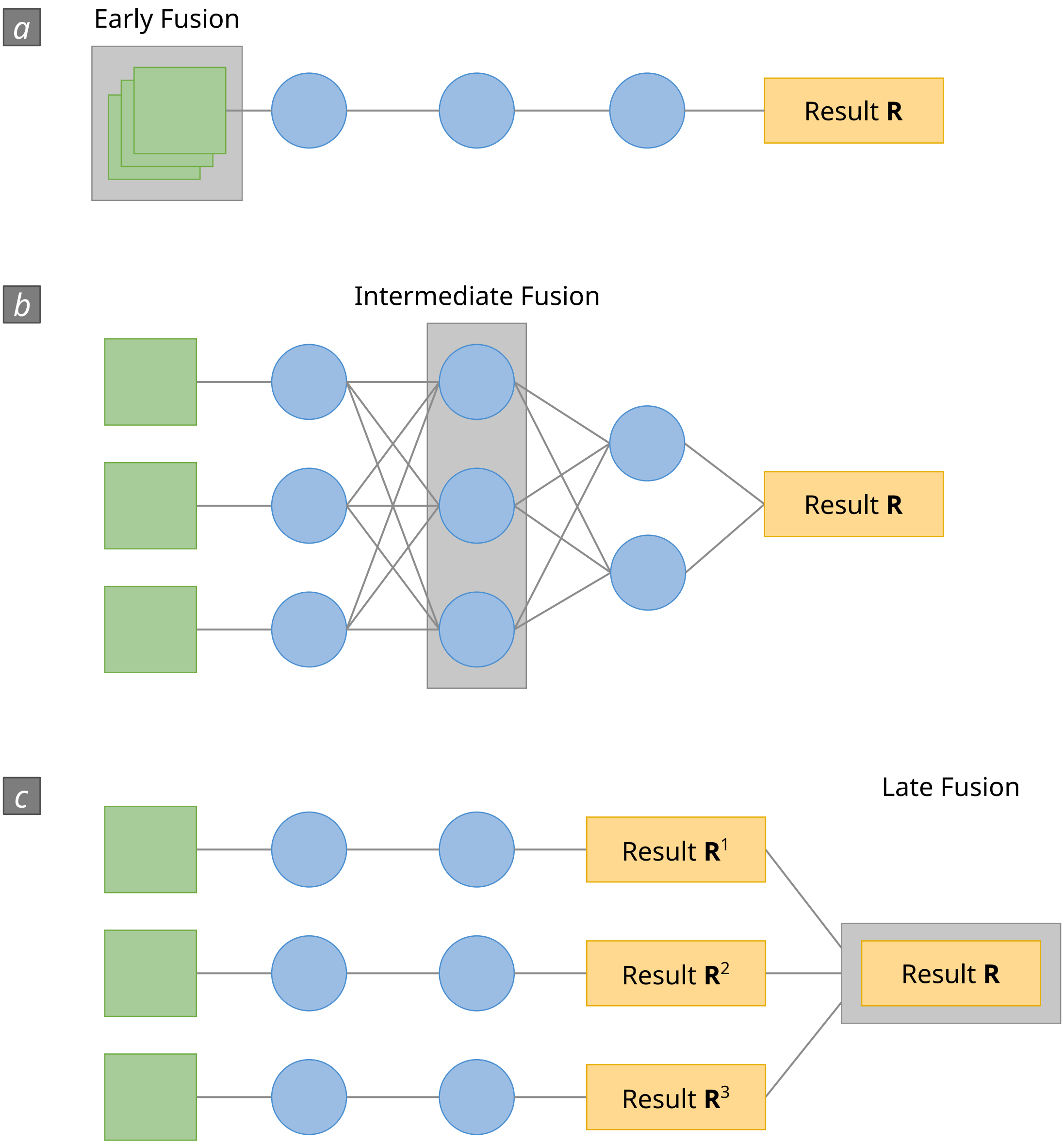


Figure 1.6: Schematic illustration of the three main fusion levels: (a) early fusion, (b) intermediate fusion, (c) late fusion. The gray boxes indicate where the fusion is performed. Figure inspired by Stahlschmidt et al. [219] and Zhou et al. [270].

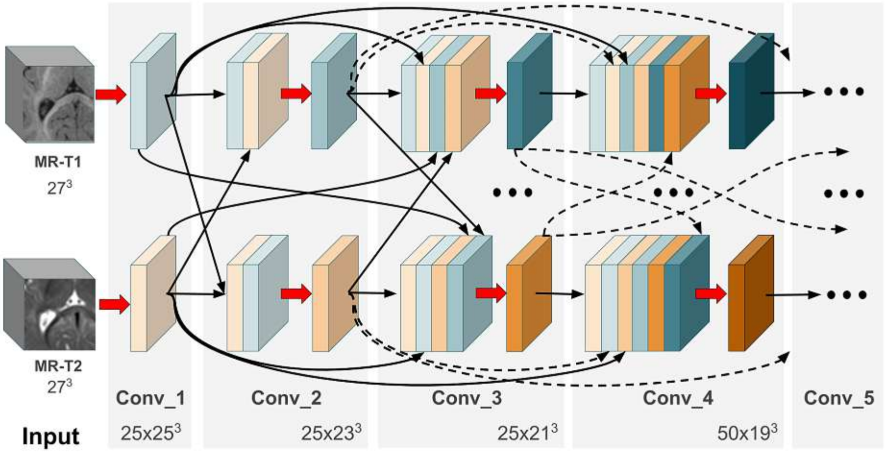


Figure 1.7: Illustration of the proposed fusion scheme in the HyperDense-Net architecture presented by Dolz et al. [57]. Image from Dolz et al. [57].

and perform early fusion, i.e., they combine all four image channels already in the input to the network. Both employ a random modality dropout strategy to reduce dependencies among modalities. Li et al. [132] use a 3D multi-scale Fully Convolutional Network (FCN) that extracts features at three different scales to better learn contextual features. They are fused at an intermediate stage later in the network for the final prediction. Das et al. [50], on the other hand, train their model on 2D images and perform simultaneous segmentation and disc identification with a region-to-image matching strategy.

There are also special cases, for example, *co-segmentation* where for the different types $\mathbf{t}^j$ of data, n type-specific results are obtained, i.e., $\mathbf{R}^j$, $j = 1, \ldots, n$, instead of only one result $\mathbf{R}$. This is, for example, used in multi-modal tumor segmentation, where anatomical and functional modalities can be combined to improve the segmentation for each of the modalities. Zhong et al. [269] utilize a 3D U-Net [43] to segment tumors in PET and CT data. In addition to the standard skip connections used in U-Net, they also introduce cross-modality connections to fuse the features from both modalities, which improves the modality-specific segmentations. In our research group, Neubauer et al. [162] propose an improved co-segmentation approach that combines early and intermediate fusion for soft-tissue tumor segmentation in combined MRI and PET/CT data. Inspired by the DenseNet architecture, they replace the standard U-Net blocks with dense blocks. This improves feature learning and reduces the number of parameters dramatically compared to Zhong et al. [269].

For multi-view data such as, for example, mammography images, the usage of *multi-view Convolutional Neural Networks (CNNs)* is actively explored in the literature. In the most general form, each view image is encoded in a separate path, also referred to as branch, and the extracted features are combined at an intermediate stage, followed by a classifier. A general multi-view architecture is illustrated in Figure 1.8.

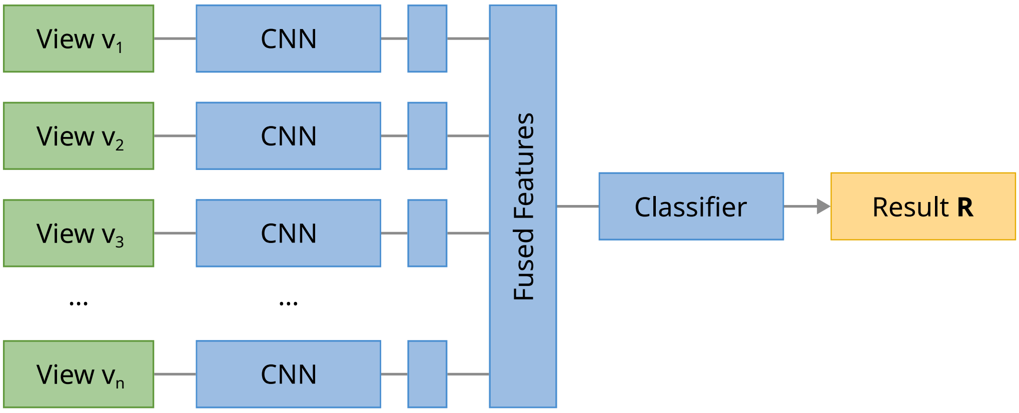


Figure 1.8: Illustration of a general deep learning-based multi-view fusion architecture.

This fusion strategy is actively employed in the field of mammography, for example, for breast density classification [102, 251] or to classify images according to their BI-RADS score [66, 30] – a standardized reporting scheme for breast imaging [212]. Carneiro et al. [30] investigate different levels of intermediate fusion and also use segmentation maps from mammographic lesions as input. Additionally, they investigate 3D inputs where they perform early fusion and treat images and corresponding segmentation masks as combined input. In all these works, different CNN architectures are being explored and used for the feature extraction part. A multi-view architecture with intermediate fusion that comprises two input branches can also be employed to detect changes between two view images/ROIs or temporal changes within one ROI [111]. Other applications of multi-view fusion are recently explored in the field of chest imaging. Hashir et al. [76] investigate the combination of posterior-anterior and lateral chest X-rays at input level as well as different intermediate levels. Further, they compare two different ways to combine the features during the fusion, namely standard feature concatenation as well as combining pixelwise statistics.

We can extract and fuse multiple views also from 3D data like chest CT scans. Setio et al. [204] propose for example false-positive reduction in pulmonary nodule detection and also utilize a multi-view CNN model for this task. They extract patches from nine different planes centered at the nodule candidate position, as shown in Figure 1.9. Each path in the multi-view CNN processes a specific view. Subsequently, multiple fusion strategies are explored by the authors. Depending on the fusion scenario, the CNN parameters may be shared or not.

Finally, we summarize a few recent methods that use an INR for representation learning from multiple views. The strength of INRs lies in how they represent features or signal values. While traditional methods discretize the input space into, e.g., voxel grids, an implicit representation is continuously defined as a "*generator function that maps input coordinates to their corresponding value within the input space*", as summarized by Molaei et al. [159]. NeRFs combine INRs with volume rendering by adding the

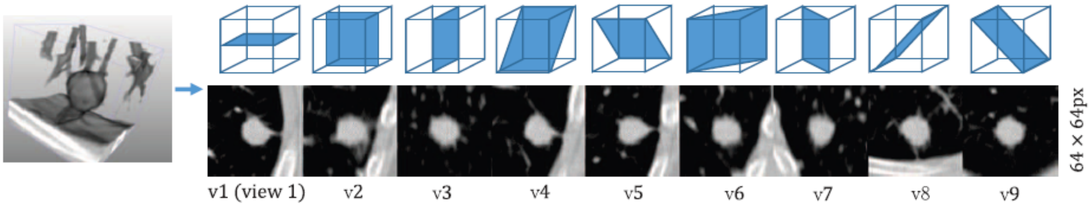


Figure 1.9: Extraction of multiple planes at a lung nodule center for false-positive reduction with a 2D multi-view CNN in the work of Setio et al. [204]. Image extracted from Figure 2 from Setio et al. [204].

viewing direction as additional input to a Multilayer Perceptron (MLP), which represents a scene or an object [157, 159]. Wu et al. [254] address the topic of high-resolution image reconstruction from low-resolution, thick-slice MR images. They model the high-resolution image as an INR, which predicts the intensity value at a given voxel position. In the reconstruction phase, the model is applied for every voxel on a dense grid, resulting in the desired high-resolution image. In a more recent work [253], they extend their method towards an arbitrary-scale super-resolution model. The authors encode voxel-specific latent representations from the low-resolution images and use them, together with the voxel positions, as input to the implicit representation network. Wu et al. [253] evaluate their method on various brain MRI datasets and also demonstrate improved segmentation performance on the superresolved data compared to other methods. For reconstruction of shapes, one method [197] uses an INR to represent the surface of the left ventricle in a cross-modality learning setup. First, they learn a shape prior from high-resolution meshes obtained from ground truth segmentation masks in cardiac CT data. Then, they apply their model to anisotropic, low-resolution left ventricle segmentations in cardiac MRI data and superresolve and complete the ventricle shape. MedNeRF [47] was proposed for reconstruction of 3D-aware CT projections from a few or just a single X-ray image. The method combines NeRFs and Generative Adversarial Networks (GANs), and additionally utilizes a self-supervised learning approach. The model captures multiple representations in a single model instead of having a separate model for different anatomical regions. The results demonstrate that MedNeRF effectively disentangles anatomy and attenuation response. A recent method [40] addresses the arbitrary-scale super-resolution problem with a NeRF-based approach. Contrary to the work by Wu et al. [253], this method does not require high-resolution data in the training, but instead learns the volumetric representation solely from a low-resolution medical volume. Instead of sampling rays, the authors propose to sample cubes at the given position, and, for the refinement of results, they apply a cube-based hierarchical rendering. The authors apply their method to slice synthesis from any viewpoint as well as to arbitrary-scale super-resolution tasks for CT and MRI data.

Late Fusion. Late fusion, i.e., decision fusion, strategies are also commonly applied for different categories of heterogeneous medical imaging data. The standard ensembling of predictions from different models, e.g., via averaging or majority voting, often also combined with data augmentation at model application time, is the simplest form of decision fusion. This is, for example, also applied in mammography in different ways [110, 154, 188]. Late fusion can be also performed with a model on top of the predictions from individual models. In the field of mammography, Kyono et al. [115] predict several radiological features (e.g., breast density, diagnosis, age) with a multi-task CNN separately for each view. In a second stage, they fuse the multi-task predictions from the four views and train a model on the predictions to obtain a benign/malignant classification for a patient.

1.5 Aim of the Work

In this thesis, we aim for methods that, on the one hand, have the potential to accelerate the image reading and diagnosis of radiologists and, on the other hand, can deal with the highly heterogeneous and complex nature of medical imaging data. Specifically, we aim for methods that fulfill the following goals:

G.1 methods that deal with changes in MRI image acquisition, i.e., with *multi-parametric* data, as well as with scans from different scanners, vendors or centers, i.e., with *multi-vendor* and *multi-center* data,

G.2 methods that are applicable *without retraining* to data from different sequences, scanners, vendors or centers, i.e., to *multi-parametric*, *multi-vendor*, and *multi-center* data,

G.3 methods that are applicable *without retraining* to data from different modalities, i.e., *multi-modal* data,

G.4 methods that combine image information from different views, i.e., *multi-view* data,

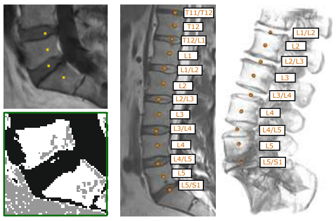
G.5 methods that combine information at different levels, e.g., patient-level, image-level, i.e., *multi-scale* data.

We addressed these goals in different research projects with our project partner and radiology software provider Agfa HealthCare [7]. Together, we identified two highly relevant use cases in clinical practice where we targeted the aforementioned data heterogeneity categories: automated *semantic spine labeling in MRI data* and *analysis of mammography imaging data*. During the course of this thesis, we developed several methods in the two clinical domains, while always following the higher-level goal of supporting and assisting radiologists in their image reading and diagnosis process.

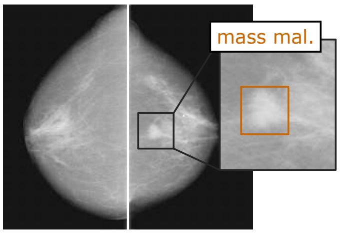
1.6 Contributions of this Thesis

The research efforts in the respective projects resulted in three main publications this thesis is based on. Chapters 2, 3, and 4 are each dedicated to the research papers that have been published at one international conference [247] and in two international journals [248, 249]. While the author of this dissertation is the first author of the presented publications, we note that they are the result of a joint effort of all involved co-authors. An overview of the papers as well as the individual contributions of all co-authors is given in the following. The three included publications are:

Paper 1 [247]: Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. "Local entropy-optimized texture models for semi-automatic spine labeling in various MRI protocols" in *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 155–159, IEEE, 2016.



Paper 2 [248]: Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. "Fully automatic cross-modality localization and labeling of vertebral bodies and intervertebral discs in 3D spinal images" in *International Journal of Computer Assisted Radiology and Surgery*, 13(10), pp. 1591–1603, 2018.



Paper 3 [249]: Maria Wimmer, Gert Sluiter, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, and Katja Bühler. "Multi-task fusion for improving mammography screening data classification" in *IEEE Transactions on Medical Imaging*, 41(4), pp. 937–950, IEEE, 2022.

We note that the respective research projects, which resulted in the three mentioned publications, were conducted from mid-2014 until 2018 (semantic spine labeling) and from 2018 until mid-2021 (analysis of mammograms). Hence, the state-of-the-art presented in Chapters 2, 3, and 4 is aligned with these periods. As this thesis was written at a later point in time, i.e., from mid-2022 until early 2023, we discuss more recent related works for both application domains in Chapter 5 of the thesis.

1.6.1 Paper 1: Local Entropy-Optimized Texture Models for Semi-Automatic Spine Labeling in Various MRI Protocols

Summary. In clinical practice, the labeled spine serves as reference for, e.g., the diagnosis of spine-related pathologies. Thus, semi- and fully automated solutions that assist the process are being actively developed. In Chapter 2, we present a novel semi-automatic pipeline for acquisition protocol independent spine labeling in volumetric MRI data of the lumbar spine from different scanners and hospitals. We specifically target the *multi-parametric* nature of MRI data and propose a method related to thesis goal **G.1**. Our learning-based system uses local three-disc ETMs, based on AAMs, for reducing the intensity scale in clinical data to only a few gray levels. The task of intervertebral disc localization is then performed on the normalized data. The benefit of this method is that we can deal with various MRI protocols, such as T1w and T2w scans, *with a single model*. Using the entropy objective allows us furthermore to apply the algorithm to acquisition protocols that are not covered by the training set, thus, also addressing goal **G.2**. We demonstrate this in an extensive evaluation on two public datasets.

Individual Contributions. The author of this thesis, *Maria Wimmer*, designed and implemented the semi-automatic labeling pipeline and the refinement method. She prepared the data and trained the three-disc models. *Maria Wimmer* conducted all experiments and performed the evaluation. She wrote the full manuscript and created all figures. *David Major* provided his expertise in spine labeling and engaged in regular technical discussions with the author of this thesis throughout the whole project. Further, he reviewed the manuscript multiple times and provided valuable feedback. *Alexey A. Novikov* participated in regular discussions and provided feedback. Further, he reviewed the manuscript several times. *Katja Bühler* proposed the initial idea of using ETMs for spine labeling in MR data. The idea to use three-disc models for spine labeling was inspired by *David Major* and *Katja Bühler* [149]. *Katja Bühler* supervised the project and the writing of the paper. She participated in regular discussions that guided the direction of the paper. Further, she provided valuable feedback and several extensive reviews of the manuscript.

Delimitation from Previous Work. The author of this dissertation conducted an initial proof-of-concept study for using ETMs [260] for spine labeling in her master's thesis [246]. One major focus of this preliminary study was to assess the mapping quality of ETMs and to find the best mapping that can be used reliably for subsequent spine labeling. The work presented in the publication "Local entropy-optimized texture models for semi-automatic spine labeling in various MRI protocols" [247] that is part of this dissertation focuses solely on spine labeling and introduces a new labeling pipeline. The two works differ in the following aspects:

- Data: In the initial study [246], one dataset comprising T1w and T2w was used, which was provided by our project partner Agfa HealthCare [7]. The work included in this thesis [247] uses two additional public datasets for training and evaluation. The new datasets are composed of T2w and MR Dixon scans from different scanners.

- Models: In the initial study, one single, global ETM was trained that covers a fixed region of the spine, i.e., T9/T10 - L5/S1. In the work by Wimmer et al. [247], several local three-disc ETMs are used. At test time, they are applied iteratively, which allows for more flexibility.

- Position refinement: In Wimmer et al. [247], a template-, non-learning-based method for refinement is presented, while in the previous study [246], probabilistic boosting trees as proposed by Schulze et al. [200] were trained and applied.

- Labeling pipeline: The usage of three-disc models and adaptive position refinement in the work by Wimmer et al. [247] required the implementation of an iterative pipeline where three-disc model matching and position refinement are performed. In the previous work [246], no labeling pipeline was used.

- Experiments: The majority of experiments in the initial study was related to finding the best mapping from source to target gray levels with ETMs. Therefore, a grid search on various ETM parameters was performed. With the best models, a position refinement method was trained and the labeling was evaluated. In contrast, the focus of Wimmer et al. [247] is solely on iterative spine labeling. Three different setups of training/testing data, including cross validation, have been evaluated.

1.6.2 Paper 2: Fully Automatic Cross-Modality Localization and Labeling of Vertebral Bodies and Intervertebral Discs in 3D Spinal Images

Summary. In Paper 2, we extend our previously presented semi-automatic ETM-based approach for anatomical labeling of the spine [247] towards a fully automatic solution. Chapter 3 introduces a cross-modality and fully automatic pipeline for labeling of intervertebral discs and vertebrae in volumetric data of the lumbar and thoracolumbar spine. The main goal of Paper 2 complies to goal **G.3** of this thesis: to provide an algorithm that is applicable not only to a wide range of multi-parametric MRI data, like T1w- and T2w MR scans, and MR Dixon data, but to multi-modal data, including also CT scans. This requires that the learned models generalize without retraining to modalities and scans with unseen image contrasts. We address this challenge by automatically localizing the sacral region combining local ETMs with CNNs. For subsequent labeling, local three-disc entropy models are matched iteratively to the spinal column, as proposed in Chapter 2. Every model-matched position is further refined by an intensity-based template-matching approach, based solely on the reduced intensity scale provided by the entropy models. We evaluate our method on a highly heterogeneous set of 161 publicly available scans, acquired on various scanners. We show that our method can deal with a wide range of different MR acquisition protocols, as well as with CT data. To the best of our knowledge, an algorithm able to deal with such a diverse set of MR and CT scans has not yet been presented in the literature by the time of this publication.

Individual Contributions. *Maria Wimmer* collected and prepared the public datasets. She designed and implemented the classifier for sacrum localization and performed the training. *Maria Wimmer* revised the labeling pipeline and implemented the template-based refinement method. Further, she conducted all experiments. This includes the seed point localization and the final labeling on MRI and CT data. She wrote the full manuscript and created all figures. *David Major* provided his expertise in spine labeling and engaged in regular technical and scientific discussions with the author of this thesis. He reviewed the manuscript several times and provided valuable feedback throughout the whole publication process. *Alexey A. Novikov* participated in regular discussions, provided feedback, and reviewed the manuscript multiple times. *Katja Bühler* supervised the project and the writing of the paper. She participated in regular discussions and provided valuable feedback that guided the direction of the paper. Furthermore, she provided multiple extensive reviews, which helped to shape the paper's story.

1.6.3 Paper 3: Multi-task Fusion for Improving Mammography Screening Data Classification

Summary. Machine learning and deep learning methods have become essential for computer-assisted prediction in medicine, with a growing number of applications also in the field of mammography. These algorithms process either one or all four views of the multi-view mammography data at once. Typically, they are trained for a *specific task*, e.g., the classification of lesions or the prediction of a mammogram's pathology status. To obtain a comprehensive view of a patient, models all trained for the *same task(s)* are subsequently ensembled or combined. In Chapter 4, we target thesis goals **G.4** and **G.5** and propose a pipeline approach where we first train a set of *individual, task-specific models* at different scales and subsequently investigate the fusion thereof, which is in contrast to standard model ensembling. Along the pipeline, we investigate different strategies to deal with the multi-view nature of mammography data. We fuse model predictions and high-level features from deep learning models with *hybrid patient meta-models* to build stronger predictors on patient level. To this end, we propose a multi-branch deep learning model that efficiently fuses features across different tasks and mammograms to obtain a comprehensive patient-level prediction. We train and evaluate our full pipeline on public mammography data, i.e., the Digital Database for Screening Mammography (DDSM) [77, 78] and its curated version, the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [121, 122]. Our experiments show that our proposed fusion approaches improve Area Under Curve (AUC) scores significantly compared to standard model ensembling. By providing not only global patient-level predictions but also task-specific model results that are related to radiological features, our pipeline aims to closely support the reading workflow of radiologists.

Individual Contributions. *Maria Wimmer* and *Katja Bühler* had the original idea of fusing the various models at different levels. *Maria Wimmer* implemented the findings classifier and the related patch-based classifier. She developed and implemented the fusion strategies and compiled the final mammography pipeline. *Maria Wimmer* and *Gert Sluiter* prepared the mammography data, and *Gert Sluiter* performed the matching of the DDSM data and prepared the data split. He implemented the view- and patient-level breast density classifier and trained the lesion localization classifier. *Maria Wimmer* and *Gert Sluiter* implemented the technical framework for the pipeline. *Maria Wimmer* conducted all experiments for the paper, i.e., she evaluated all task-specific models as well as all fusion approaches. Further, she conducted all ablation studies, i.e., retrained and evaluated the findings and breast density classifiers as well as the fusion models. *Maria Wimmer* wrote the full manuscript and created all figures. *David Major*, *Dimitrios Lenis*, and *Astrid Berg* provided technical input regarding the fusion strategies. *David Major* engaged in regular technical discussions and gave feedback. He reviewed the manuscript in detail several times throughout the whole publication process. *Dimitrios Lenis* participated in regular technical discussions and contributed his expertise in statistical testing. *Dimitrios Lenis*, *Astrid Berg*, and *Theresa Neubauer* thoroughly reviewed the manuscript. *All co-authors* participated in regular discussions and gave feedback. *Katja Bühler* supervised and coordinated the project and the writing of the paper. She provided regular input and feedback and reviewed the manuscript multiple times, which helped to shape the paper's story and focus.

1.6.4 Related Co-Authored Publications

During the course of this thesis, the author was involved in different projects related to the topic and selected application domains of this thesis. These projects resulted in various scientific publications that (a) aim to assist and improve radiologists' image interpretation and diagnosis tasks and (b) propose methods that address different categories of heterogeneous data with standard, generalization- and fusion-based approaches.

A list of the selected publications is given below. In three papers [164, 165, 267], we follow the standard approach and propose methods tailored to a specific type (and body region). In Zheng et al. [267], we introduce a spine labeling and segmentation method specific to T2w data for a computational challenge at MICCAI 2015. Novikov et al. [164, 165] present two deep learning-based segmentation approaches: one is a 2D segmentation method for lung X-rays [164], the second one a sequential segmentation approach trained and evaluated on liver and vertebrae CT data [165]. The paper by Neubauer et al. [162] proposes a fusion method for multi-modal co-segmentation of soft-tissue sarcomas in combined MRI and PET/CT data, as already described in Section 1.4.2. Finally, the works by Major and Lenis are concerned with domain awareness in medical image classifier interpretation [125, 150, 151]. Their approaches aim for more precise and faithful classifier decision visualization as compared to currently used methodology. Their methods are evaluated on mammography and chest X-ray data. In the most recent work [150], they use an image classifier introduced in Wimmer et al. [249] (see Chapter 4) that generalizes across the different mammography views.

The following mentioned co-authored publications are not part of this thesis:

- David Major, Dimitrios Lenis, **Maria Wimmer**, Astrid Berg, Theresa Neubauer, Katja Bühler. "On the importance of domain awareness in classifier interpretations in medical imaging" in *IEEE Transactions on Medical Imaging*, 42(8), pp. 2286–2298, IEEE, 2023.

- Theresa Neubauer, **Maria Wimmer**, Astrid Berg, David Major, Dimitrios Lenis, Thomas Beyer, Jelena Saponjski, and Katja Bühler. "Soft Tissue Sarcoma Co-segmentation in Combined MRI and PET/CT Data" in *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (ML-CDS/CLIP 2020)*, vol. 12445 of Lecture Notes in Computer Science, pp. 97–105, 2020.

- Dimitrios Lenis, David Major, **Maria Wimmer**, Astrid Berg, Gert Sluiter, and Katja Bühler. "Domain aware medical image classifier interpretation by counterfactual impact analysis" in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, vol. 12261 of Lecture Notes in Computer Science, pp. 315–325, 2020.

- David Major*, Dimitrios Lenis*, **Maria Wimmer**, Gert Sluiter, Astrid Berg, and Katja Bühler. "Interpreting Medical Image Classifiers by Optimization Based Counterfactual Impact Analysis" in *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1096–1100, IEEE, 2020. * equal contribution.

- Alexey A. Novikov, David Major, **Maria Wimmer**, Dimitrios Lenis, and Katja Bühler. "Deep Sequential Segmentation of Organs in Volumetric Medical Scans" in *IEEE Transactions on Medical Imaging*, 38(5), pp. 1207–1215, IEEE, 2019.

- Alexey A. Novikov, Dimitrios Lenis, David Major, Jiří Hladůvka, **Maria Wimmer**, and Katja Bühler. "Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs" in *IEEE Transactions on Medical Imaging*, 37(8), pp. 1865–1876, IEEE, 2018.

- Guoyan Zheng, Chengwen Chu, Daniel L. Belavý, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lôpez Andrade, Ben Glocker, Hao Chen, Qi Dou, Pheng-Ann Heng, Chunliang Wang, Daniel Forsberg, Aleŝ Neubert, Jürgen Fripp, Martin Urschler, Darko Stern, **Maria Wimmer**, Alexey A. Novikov, Hui Cheng, Gabriele Armbrecht, Dieter Felsenberg, Shuo Li. "Evaluation and Comparison of 3D Intervertebral Disc Localization and Segmentation Methods for 3D T2 MRI Data: A Grand Challenge" in *Medical Image Analysis*, vol. 35, pp. 327–344, 2017.

CHAPTER 2

Semi-Automatic Spine Labeling in Multi-Sequence MRI Data

This chapter is based on the following publications:

- Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. "Local entropy-optimized texture models for semi-automatic spine labeling in various MRI protocols" in *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 155–159, IEEE, 2016. DOI: 10.1109/ISBI.2016.7493233
- Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. "Fully automatic cross-modality localization and labeling of vertebral bodies and intervertebral discs in 3D spinal images" in *International Journal of Computer Assisted Radiology and Surgery*, 13(10), pp. 1591–1603, 2018. DOI: 10.1007/s11548-018-1818-3

2.1 Introduction

Labeling of the spinal column is a time-consuming, yet necessary task for clinicians as it serves as anatomical reference system to describe the location of organs and diagnostic findings. In spinal imaging, it serves in both diagnosis and preoperative planning of spine-related injuries and pathologies. The term *"labeling"* refers in this context to the assignment of an *anatomical label* to a landmark position within spinal tissue. In the context of this work, the landmark position corresponds to the center of a vertebra or intervertebral disc, while the label denotes the anatomical name of the corresponding tissue. The human vertebral column comprises *7 cervical*, *12 thoracic*, and *5 lumbar vertebrae* that are separated by *intervertebral discs*, as depicted in Figure 2.1. The spine consists of 23 intervertebral discs in total, with no disc between the two topmost cervical vertebrae. Below the lumbar spine are two bony structures of fused vertebrae, the *sacrum*, comprising 5 fused vertebrae, and the *coccyx*.

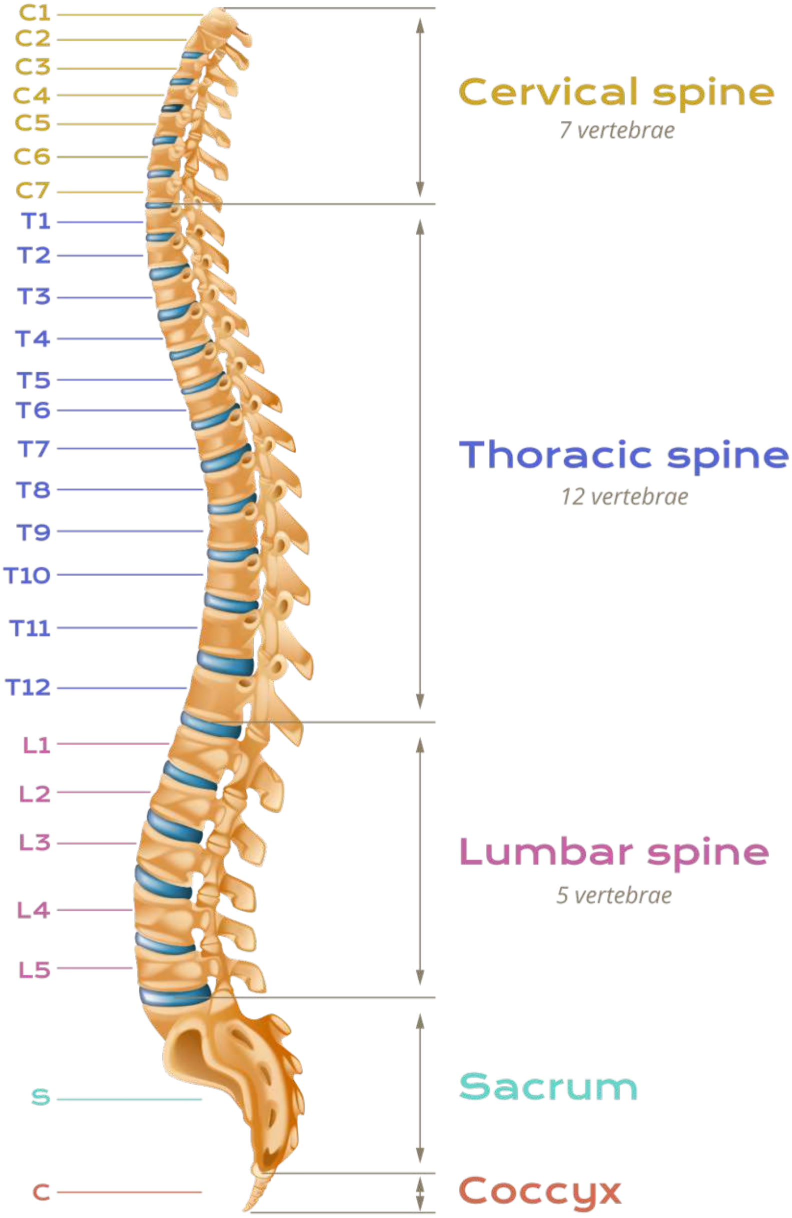


Figure 2.1: Anatomical illustration of the human vertebral column.
Adapted from image designed by Macrovector - Freepik.com (Accessed: November 12, 2022).

In clinical practice, MRI and CT imaging data is widely used to analyze the spine [184]. The growing number of scans that are being acquired has necessitated semi- and fully automated solutions to accelerate the respective radiological workflows. However, developing computer-aided tools for spine labeling is challenging, especially for applicability to MR data. MR scans lack a generalized intensity scale like Hounsfield Units in CT [184]. Changing the scan parameters results in different image contrasts, and a wide range of MR sequences and acquisition protocols that are nowadays available. Tissue intensities are further influenced by imaging artifacts and the use of scanners from different vendors. Hence, the same tissue, e.g., vertebrae, can have varying appearances even within a single scan. Physiological and pathological changes in the body further affect the appearance [241].

2.1.1 Related Work

The following literature review addresses related spine labeling papers published prior to the work by Wimmer et al. [247], on which this chapter is based. We refer to Section 5.2 for a discussion of more recent works in that field.

Various recent works target spine labeling in *specific* MR sequences, e.g., on T1w [156], T2w [36, 48], Dixon [35] or SPIR [12] data. All these methods are specific to a certain protocol. Thus, approaches which are able to localize the spinal parts without retraining for the different imaging parameters are of high interest. Lootus et al. [141] apply Deformable Part Models using the Histogram of Oriented Gradients descriptor combined with a graphical model to localize vertebrae in a set of T2w MRI scans. The authors claim that their method is applicable without retraining also on CT data, although no extensive evaluation is provided. In a more recent work [142], they introduce a normalization based on the median vertebral intensities. This makes their algorithm more robust to parameter changes and different vendors in T2w MRI scans as compared to their previous approach [141]. Zukić et al. [272] introduce an automated detection and segmentation framework for vertebrae using a boosted cascade of simple features for detection and the watershed method for segmentation. Their method is able to handle T1w, Turbo Inversion Recovery Magnitude (TIRM), and T2w MRI data at once, although the extension to other sequence types requires a retraining of their detectors. Cai et al. [28] address the problem of labeling and segmenting the spine in a modality independent way. A combination of standard and convolutional restricted Boltzmann machine layers is used for landmark detection, followed by a global spine model matching algorithm. Finally, vertebrae are registered and segmented via local models. Their framework is evaluated on MR and CT data. It is applicable without retraining to different modalities whereby the modality information of the unseen scan is required at model test time. However, it is much more complex than our proposed method.

2.1.2 Contribution

We present a novel *semi-automatic pipeline* for labeling of the spinal column, that can process 3D MR scans with a high intensity variability, like T1w and T2w scans, acquired on different scanners with varying scan parameters. These properties relate directly to goal **G.1** of this thesis. We propose a *learning-based system*, where we train *local Entropy-Optimized Texture Models (ETMs)* [260] for reducing the intensity scale in the clinical data. We build general models that normalize data across different MR protocols with the goal of finding rather homogeneous mappings for intervertebral discs, vertebrae, and the spinal cord. Thus, no separate models for different acquisition setups are required. Moreover, our method is applicable to sequences and protocols which are not covered by the training set, thus, also addressing thesis goal **G.2** in this chapter. When labeling an unseen scan, the learned models are applied and intervertebral disc centers are localized in an iterative labeling pipeline. The disc centers are further refined with an *adaptive position refinement* to increase the disc center accuracy. This method does not require any training and is based solely on the normalized data.

2.2 Methods and Materials

The following section first revisits the general concept of ETMs as proposed by Zambal et al. [260] in Section 2.2.1. We then present our proposed application of ETMs within the scope of spine labeling in Section 2.2.2. Section 2.2.3 describes the final labeling pipeline, including the suggested position refinement.

Notation: We define a volume $V \subset \mathbb{R}^3$ as the sagittal 3D reconstruction of a scan, i.e., a stack of 2D slices, where the xy-plane is the sagittal plane, and z the medial-lateral depth. Further, we define a set of volumes $\mathcal{D}_{tr}$ with $V_k \in \mathcal{D}_{tr}, k = 1, \dots |\mathcal{D}_{tr}|$ for training of models and an unseen test set $\mathcal{D}_{test}$, with $V_j \in \mathcal{D}_{test}, j = 1, \dots |\mathcal{D}_{test}|$. A position within V is denoted by $\boldsymbol{p} = (x, y, z)$. Further, we define a set of anatomical labels Λ, including labels for vertebrae and intervertebral discs, following a standard anatomical atlas of the human spine [8] as: $\Lambda = (S1, L5/S1, L5, L4/L5, L4, \dots, C3, C2/C3)$, where λ_l corresponds to the label at the l-th position in the ordered labelset Λ, $l = 1, \dots, |\Lambda|$. Here, $S1$ refers to the first vertebra of the sacrum, L denotes lumbar, T thoracic, and C cervical vertebrae (see Figure 2.1). Consequently, an intervertebral disc between two vertebrae, e.g., between $L4$ and $L5$, is called $L4/L5$. $C2/C3$ is the topmost intervertebral disc, as no disc exists between $C1$ and $C2$. In this thesis we do not consider the remaining four vertebrae of the sacrum and the coccyx. Therefore, they are excluded from Λ. We refer to a position labeled with a certain anatomical label $\lambda_l \in \Lambda$ as $\boldsymbol{p}^{\lambda_l} \in V$. $\tilde{\boldsymbol{p}}$ denotes a model-matched position, i.e., the position of a model landmark in a scan after matching a model to the scan. Further, we refer to a position after applying a refinement method with $\hat{\boldsymbol{p}}$. Finally, the annotated center of an intervertebral disc or vertebra in a scan, i.e., the ground truth position, is denoted with $\boldsymbol{d}^{\lambda_l}$.

2.2.1 Entropy Texture Models in General

Zambal et al. [260] introduced Entropy-Optimized Texture Models (ETMs) for segmentation from dense landmarks at object boundaries in 2D medical imaging data. They propose a novel texture model, that extends Active Appearance Models (AAMs) with a texture normalization approach, as already briefly mentioned in Section 1.4.1. Instead of texture modeling based on Principal Component Analysis, the model texture is described by probability density functions of a reduced set of target gray values. The related mappings of the input gray levels to the target levels are optimized in terms of entropy, hence, minimizing the uncertainty of mappings while maximizing information content. The idea of using entropy is borrowed from multi-modal image registration, where the entropy-based mutual information criterion is used. The novel texture representation as well as using the entropy objective are powerful features of ETMs that allow building of robust models against texture variations, like brightness variations, and even different imaging modalities.

Training. To train an entropy model M, we first require a set of corresponding landmarks in each $V_k, k = 1, \ldots, |\mathcal{D}_{tr}|$ where $\mathcal{D}_{tr}$ refers to the training set. Training textures $T_k \subset V_k$ are extracted from every volume V_k in the following way. According to Zambal et al. [260], the 3D convex hull defined by the landmarks is consistently tetrahedralized and the texture T_k is extracted and resampled by N texels, similar to AAMs. The term texel formally refers to a pixel in a texture. All T_k are initially quantized to r source gray levels; hence, every texture is in the same intensity range. The task is to find optimal mappings f_k for every texture T_k, that map the r source levels to a reduced number of t target gray levels. The mappings are found with iterative optimization of two entropy-based objectives H^{model} and H^{tex} [260]. The *model entropy* H^{model} describes the uncertainty of all mappings and is minimized:

$$H^{model} = \frac{1}{N} \sum_{j=1}^{N} H(p_j) \ \rightarrow \min \tag{2.1}$$

where p_j denotes the Probability Density Function (PDF) of observed target gray levels across all mapped training textures at the j-th texel, and $H(p_j)$ the entropy of the PDF. The minimization of H^{model} ensures that each texel maps to a certain target gray value. However, H^{model} yields a minimum when all texels map to the same target gray value.

The *image entropy* H^{tex} compensates for that and prevents the degeneration of mappings f_k, i.e., so that not all mappings map to the same target value:

$$H^{tex} = \frac{1}{|\mathcal{D}_{tr}|} \sum_{k=1}^{|\mathcal{D}_{tr}|} H(f_k(T_k)) \ \rightarrow \max \tag{2.2}$$

This term drives the mappings to maximal information content. Simulated annealing is used to find optimal mappings f_k in the training, subject to maximizing $(H^{tex} - H^{model})$.

In summary, H^{model} ensures, that the same tissues are normalized to the same target level, while H^{tex} guarantees that the contrast between different tissues is preserved. This enables us to apply such a model to images that exhibit similar intra-tissue homogeneity and inter-tissue contrast as those images captured by the model during training. The output of the ETM training phase is the local model M, which captures the variation of the shape and texture of the underlying anatomy.

Model Matching. Matching M to an unseen scan, we iteratively change its shape and texture mapping and assess the matching quality. This is done based on the model PDFs and texture U currently overlapped by the model in the following way. In every iteration, we change the shape within the valid shape space defined by the mean shape and a linear combination of the shape eigenvectors [46]. Next, we extract texture U currently overlapped by the model as described in the model training phase. Further, we obtain an intensity mapping f_u for U by assigning the target values for every texel j, $j = 1, \ldots, N$, that lead to the maximum likelihood according to the texture model PDFs p_j. We assess the matching quality in every iteration via Bayesian reasoning using the naive Bayesian assumption by maximizing the posterior probability of the current shape given the normalized texture U' [260]. Finally, the obtained mapping f_u is applied. This results in the desired intensity-reduced scan with only t target gray levels.

2.2.2 ETMs for Spine Labeling

For spine labeling, we utilize the capabilities of ETMs for data normalization and propose to build *local three-disc models* M^{λ_l} from a mixed set of T1w and T2w MR volumes (see Figure 2.2). The use of three-disc models for spine labeling has been already successfully applied in the literature, for example, by Major et al. [149].

We build the models M^{λ_l} from sparse landmarks around an annotated *middle disc* $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$ including also its adjacent *upper disc* $\boldsymbol{d}^{\boldsymbol{\lambda_{l+2}}}$ and *lower disc* $\boldsymbol{d}^{\boldsymbol{\lambda_{l-2}}}$. Figure 2.3 illustrates the extracted landmarks. Around each intervertebral disc center, we include sampled positions along the surface of a cylinder, which approximates the size of the disc. Furthermore, we add the center positions $\boldsymbol{d}^{\boldsymbol{\lambda_{l+1}}}$ and $\boldsymbol{d}^{\boldsymbol{\lambda_{l-1}}}$ with labels λ_{l+1} and λ_{l-1}, respectively, of the two vertebrae that lie between the upper disc $\boldsymbol{d}^{\boldsymbol{\lambda_{l+2}}}$ and lower disc $\boldsymbol{d}^{\boldsymbol{\lambda_{l-2}}}$. The vertebrae centers $\boldsymbol{d}^{\boldsymbol{\lambda_{l+1}}}$ and $\boldsymbol{d}^{\boldsymbol{\lambda_{l-1}}}$ indicate the centers of the respective vertebral bodies. Finally, we also add landmarks in the spinal canal that correspond to the disc and vertebrae centers (see Figure 2.3).

We extract this set of landmarks from different T1w and T2w MR volumes V_k in the training set $\mathcal{D}_{tr}$ and train model M^{λ_l} according to the steps described in Section 2.2.1 (see Figure 2.2). We repeat this process for various middle discs $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$, resulting in a set of different three-disc models M^{λ_l} which we combine for labeling an unseen scan.

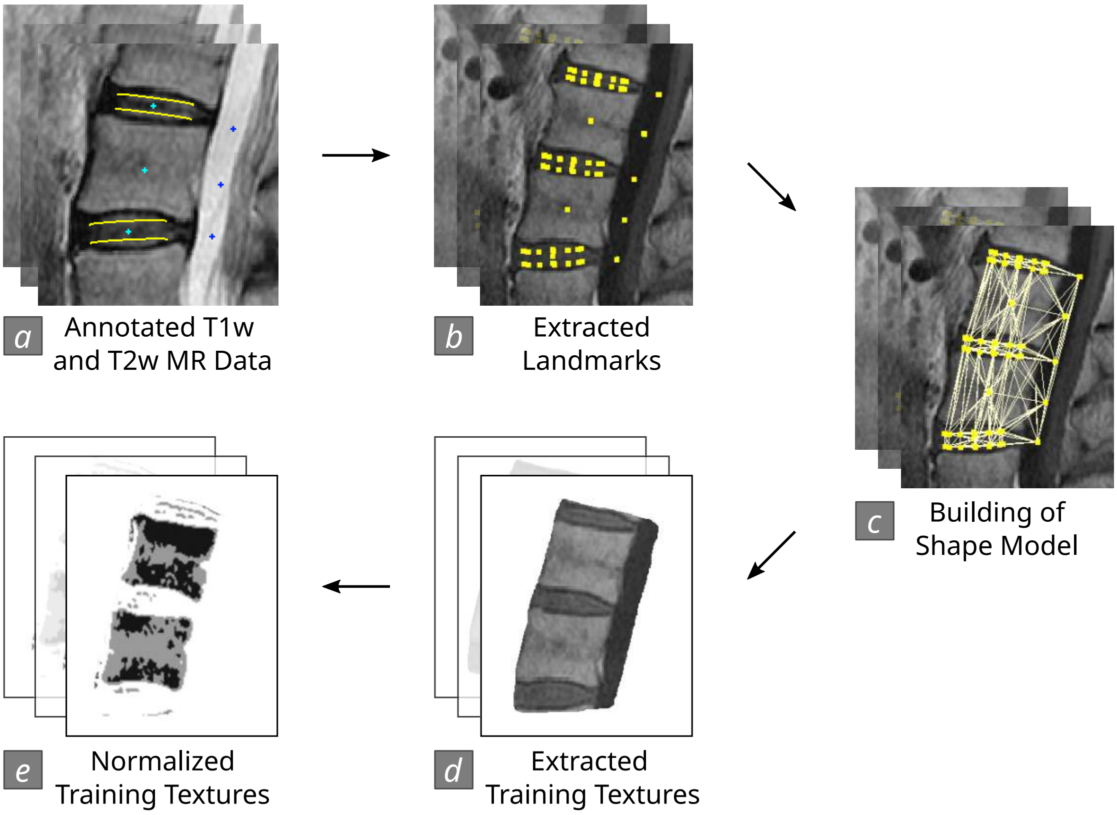

a
Annotated T1w and T2w MR Data
b
Extracted Landmarks
c
Building of Shape Model
d
Extracted Training Textures
e
Normalized Training Textures

Figure 2.2: Overview of the training of a three-disc ETM M^{λ_l} for data normalization: (a) From an annotated set of MR data with: intervertebral disc and vertebrae centers (blue), spinal canal landmarks (blue), cylinders that approximate the intervertebral discs (yellow), (b) corresponding model landmarks are extracted (yellow), (c) and a shape model is built. (d) Training textures are extracted and the texture transformations are then optimized iteratively [260]. (e) The results are normalized training textures and the trained three-disc model M^{λ_l}.

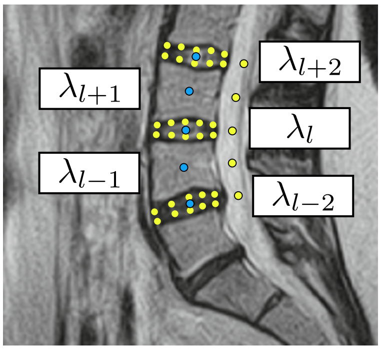


Figure 2.3: Schematic illustration of landmarks extracted for three-disc models M^{λ_l} for labeling. Intervertebral disc and vertebrae center positions are depicted in blue, remaining model landmark positions in yellow.

2.2.3 Labeling of an Unseen Volumetric Scan

The proposed labeling pipeline is illustrated in Figure 2.4. Our semi-automatic method requires minimal input from a user for labeling an unseen scan: an initial click position $\boldsymbol{p}$ in volume V inside an intervertebral disc or vertebra and its corresponding label λ_l, i.e., we require position $\boldsymbol{p}^{\lambda_l}$. This position serves as initialization for the iterative labeling, which repeats the following steps: ETM matching, adaptive disc center position refinement, and propagation.

ETM Matching. We initialize the learned three-disc model M^{λ_l} that corresponds to the anatomical label λ_l at position $\boldsymbol{p}^{\lambda_l}$ and match it in an iterative manner, as described in Section 2.2.1. We retrieve the normalized scan V' as well as the model-matched positions, denoted as $\tilde{\boldsymbol{p}}$. Specifically, we consider the intervertebral disc *candidate positions*, i.e., the set of model-matched labeled disc positions $\{\tilde{\boldsymbol{p}}^{\lambda_{l+2}}, \tilde{\boldsymbol{p}}^{\lambda_l}, \tilde{\boldsymbol{p}}^{\lambda_{l-2}}\}$ for the upper disc, middle disc, and lower disc, respectively (see Figure 2.4 (c)). These positions should lie already in the vicinity of the respective ground truth centers $\boldsymbol{d}^{\boldsymbol{\lambda_{l+2}}}$, $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$, and $\boldsymbol{d}^{\boldsymbol{\lambda_{l-2}}}$. However, to account for imperfections in the model matching, we apply a refinement step for the candidate disc center position of the middle disc $\tilde{\boldsymbol{p}}^{\lambda_l}$.

Adaptive Disc Center Position Refinement. The main objective of this step is to further increase the disc center position accuracy of $\tilde{\boldsymbol{p}}^{\lambda_l}$. Instead of training a learning-based method, we propose an adaptive approach inspired by *Haar-like features* [233] which solely uses the normalized data V' and model-matched positions $\tilde{\boldsymbol{p}}$. We span a bounding box around the model-matched disc position $\tilde{\boldsymbol{p}}^{\lambda_l}$, defining our search region $\mathcal{R}$ for refinement (see Figure 2.5 in orange). The size of $\mathcal{R}$ is based on the ground truth, from which we calculate the average dimension of discs in sagittal, axial, and coronal direction. For every voxel inside $\mathcal{R}$ we decide if it belongs to the disc by applying a filter inspired by Haar-like features [233]. We construct the filter with three regions: upper region $\mathcal{R}_U$, middle region $\mathcal{R}_M$, and lower region $\mathcal{R}_L$, and place them in following way: $\mathcal{R}_M$ is centered at the current voxel position $\boldsymbol{p}'$ in $\mathcal{R}$. $\mathcal{R}_U$ and $\mathcal{R}_L$ are displaced based on the intervertebral disc orientation vector $\boldsymbol{n}$ and the average disc thickness estimated from the ground truth. Vector $\boldsymbol{n}$ is calculated based on the model-matched landmark positions. For every region $\mathcal{R}_U$, $\mathcal{R}_M$ and $\mathcal{R}_L$, we calculate its most frequent intensity value, i.e., the *intensity mode*: m_L, m_M and m_U. We consider the voxel $\boldsymbol{p}'$ as potential disc position if:

$$Mask(x, y, z) = \begin{cases} 1 & \text{if } m_U \neq m_M \wedge m_L \neq m_M \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

From the obtained binary mask, we calculate the *centroid* as the refined center position $\hat{\boldsymbol{p}}^{\lambda_l}$ for the disc with label λ_l.

Propagation. The labeling is performed in an iterative manner. From the model matched at initial click position $\boldsymbol{p}^{\lambda_l}$ we also obtain candidate positions for the upper and

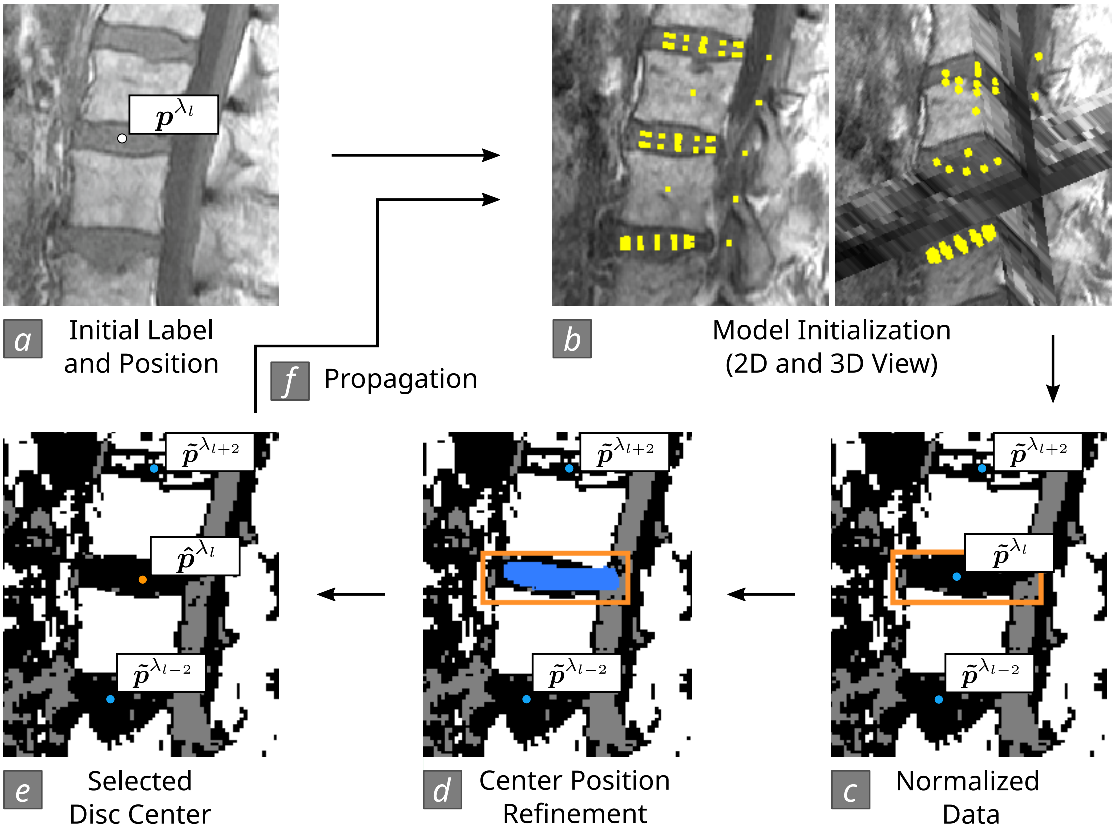

p^{λ_l}
a Initial Label and Position
b Model Initialization (2D and 3D View)
f Propagation
$\tilde{p}^{\lambda_{l+2}}$
$\hat{p}^{\lambda_l}$
$\tilde{p}^{\lambda_l}$
$\tilde{p}^{\lambda_{l-2}}$
e Selected Disc Center
d Center Position Refinement
c Normalized Data

Figure 2.4: Basic steps when labeling a 3D MR scan: (a) Based on the initial position $\boldsymbol{p}^{\lambda_l}$ with corresponding label λ_l provided by a user, (b) the model M^{λ_l} is placed in the scan and matched. (c) The results are the normalized scan V' and the model-matched positions $\tilde{\boldsymbol{p}}$, which serve as input to (d) the adaptive refinement method. By applying a filter within a search region (orange) around position $\tilde{\boldsymbol{p}}^{\lambda_l}$, we obtain a binary mask (blue). (e) The refined center position $\hat{\boldsymbol{p}}^{\lambda_l}$ with label λ_l is the centroid of the mask. (f) We place the next three-disc model at position $\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$ or $\tilde{\boldsymbol{p}}^{\lambda_{l+2}}$, depending on the search direction, and continue from (b).

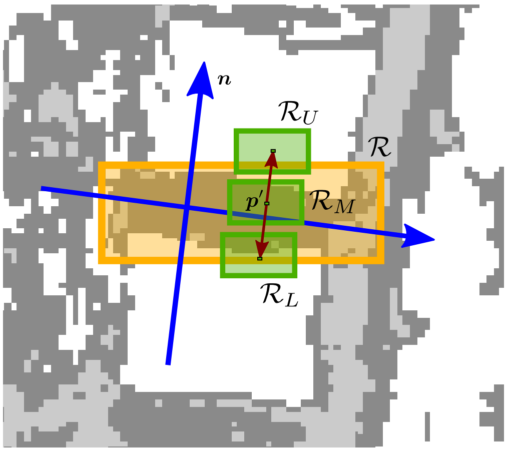


Figure 2.5: Illustration of our proposed refinement method. We construct a filter comprising the regions $\mathcal{R}_U$, $\mathcal{R}_M$, and $\mathcal{R}_L$ (green), which are displaced according to the intervertebral disc orientation. Displacement vectors are shown in red and the intervertebral disc orientation vectors in blue. We match the filter at each voxel in the search region $\mathcal{R}$ (orange) and decide if the voxel belongs to the disc or not.

lower disc $\tilde{\boldsymbol{p}}^{\lambda_{l+2}}$ and $\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$, respectively (see Figure 2.4 (e)). We continue the search first towards $L5/S1$ and place the next three-disc model $M^{\lambda_{l-2}}$ at $\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$. We repeat model matching, adaptive refinement, and retrieval of the next intervertebral disc position $\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$ until we reach $L5/S1$. Then, we continue the search upwards to the topmost disc $C2/C3$, starting from the initial click position $\boldsymbol{p}^{\lambda_l}$ respectively the model-matched position $\tilde{\boldsymbol{p}}^{\lambda_{l+2}}$.

2.3 Evaluation and Results

The following section provides an elaborate evaluation of our novel pipeline on lumbar MR volumes.

2.3.1 Data and Experimental Setup

Four datasets $\mathcal{D}_i$ of sagittally acquired lumbar MR scans from five different scanners are used for evaluation. The voxel sizes are highly anisotropic, with an in-plane resolution ranging from 0.59 mm^2 to 1.25 mm^2 and slice thickness from 2 to 6 mm. We reconstruct 62 volumes from the scans and consider a subset of seven intervertebral discs covering the region from $L5/S1$ to $T11/T12$, which results in 434 discs in total. The four datasets $\mathcal{D}_i$ comprise the following scans:

- **Dataset $\mathcal{D}_1$:** Our private dataset with 13 T1w MR scans.
- **Dataset $\mathcal{D}_2$:** Our private dataset with 10 T2w MR scans.

- **Dataset $\mathcal{D}_3$:** Challenge dataset [36], comprising T2w MR scans from 15 different subjects.

- **Dataset $\mathcal{D}_4$:** MR Dixon data [35]. Scans from eight subjects, where we use the following three image channels: opposed-phase, fat, and water saturated image. We treat every channel separately, hence we obtain 24 volumes for testing.

In our private datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, magnetic field inhomogeneities are present and half of the scans exhibit at least one of the following pathologies: fractures, disc herniation, scoliosis, and lumbar hyperlordosis. No details about present pathologies are provided for datasets $\mathcal{D}_3$ and $\mathcal{D}_4$.

We evaluate three different setups as summarized in Table 2.1. For the first setup #1, we split $\mathcal{D}_1$ and $\mathcal{D}_2$ into two subsets and perform two-fold-cross validation. We refer to the cross-validation subsets of $\mathcal{D}_1$ with $\mathcal{D}_{1,1}$ and $\mathcal{D}_{1,2}$, and analogously also for $\mathcal{D}_2$. Setups #2 and #3 demonstrate the generality of our method by evaluation on unseen MR sequences and image contrasts, which are *not included* in the training set. All experiments are conducted on an Intel Xeon E5 PC.

Setup	Training Data $\mathcal{D}_{tr}$		Test Data $\mathcal{D}_{test}$		
#1	$\mathcal{D}_{1,1}$	$\mathcal{D}_{2,1}$	$\mathcal{D}_{1,2}$	$\mathcal{D}_{2,2}$	
	$\mathcal{D}_{1,2}$	$\mathcal{D}_{2,2}$	$\mathcal{D}_{1,1}$	$\mathcal{D}_{2,1}$	
#2	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	
#3	$\mathcal{D}_3$		$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_4$

Table 2.1: Evaluation setups.

2.3.2 ETM Model Training

We train three-disc models M^{λ_l} for each middle disc from $L4/L5$ up to $T12/L1$. This introduces an overlap in the region that is covered by the models and increases robustness. Choosing parameters for ETMs, the goal is to find rather homogeneous intensity mappings for intervertebral discs, vertebrae, and the spinal cord, without losing anatomical information. Best mapping results are obtained with $r = 110$ source and $s = 3$ target levels using leave-one-out cross-validation on our training data. This provides a good initial quantization where we do not miss relevant intensity changes and at the same time remove image noise. The time for model training depends on the number of training textures. We report 4.6 ± 1.5 min for setups #1 and #3 and 13 ± 2.3 min for setup #2 on average.

It is important to note, that we do not obtain the same mapping for one tissue in all sequences, e.g., for the mapping of disc intensities. The tissues are in a different range, but still homogeneous mappings are obtained. Sample images are depicted in Figure 2.6.

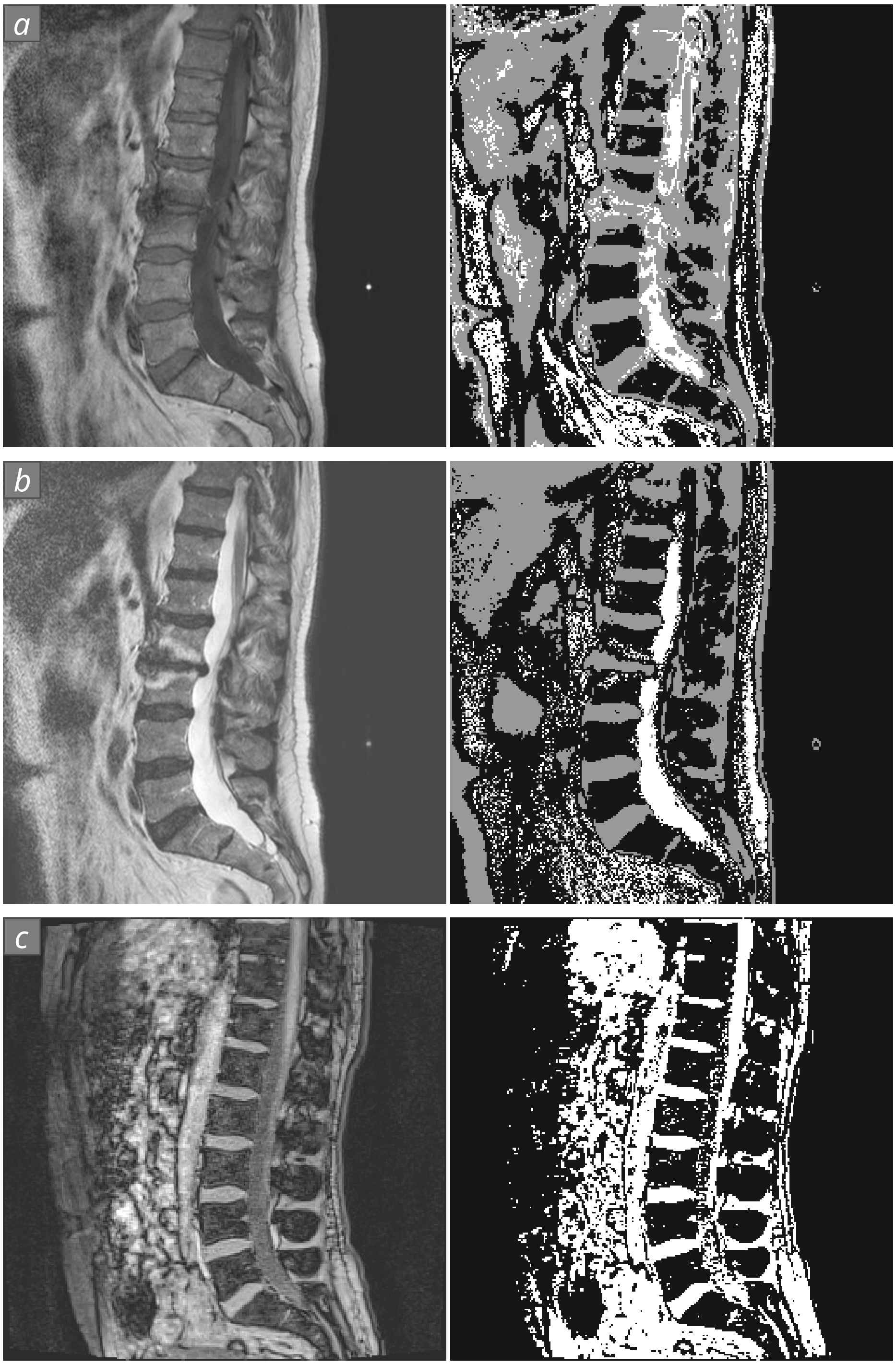


Figure 2.6: Various mid-sagittal slices of MR scans from our datasets $\mathcal{D}_i$ (left) and corresponding sample normalization results (right): (a) T1w, (b) T2w, and (c) Dixon opposed-phase image channel.

2.3.3 Labeling Results

We start a full semi-automatic data normalization and labeling run inside every intervertebral disc. Initial seed positions are sampled within 10 mm to the ground truth center for our experiments. This allows the models to converge towards the disc centers while matching. The overall processing time for labeling an unseen MR scan is 12.6 ± 3.7 seconds, which results in 1.8 ± 0.5 seconds per intervertebral disc.

To measure the performance of our method, we calculate the following two evaluation metrics that have been introduced at the intervertebral disc localization challenge at MICCAI 2015 [267]:

- p_ϵ – **intervertebral disc localization accuracy:** p_ϵ explains, how many detected positions $\hat{\boldsymbol{p}}^{\lambda_l}$ lie within ϵ millimeter to the corresponding annotated disc center $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$, with $\epsilon \in \{2, 4, 6, 10\}$.
- $\bar{e} \pm sd$ – **intervertebral disc center accuracy:** The Euclidean distance from the localized position $\hat{\boldsymbol{p}}^{\lambda_l}$ to the ground truth $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$ describes the mean position error $\bar{e}$ and standard deviation sd.

Labeling results are shown in Figure 2.7. The highest intervertebral disc center position accuracy is achieved with setup #2 (see Figure 2.7 (a)), where we reach a mean error of 3.82 ± 2.47 mm with $p_{10} = 97.64\%$ (see Table 2.2). A high localization rate is also achieved for setup #3 with $p_{10} = 92.06\%$ and an average distance of 4.45 ± 3.44 mm from the ground truth. The overall better performance of setup #2 compared to setup #3 can be explained by the higher diversity of the training data and the larger training set in setup #2. Setups #2 and #3 are trained on only a *subset* of all available MR sequences and protocols in $\mathcal{D}$ and demonstrate the generalization capabilities of our method. Thus, we show that our pipeline generalizes to acquisition protocols that are not included in the training, but exhibit similar contrast characteristics as captured by our learned ETMs. This is reflected in the results on $\mathcal{D}_4$, which was never included in the training process. In general, we observe higher p_ϵ for T2w scans as compared to T1w scans, because of a better tissue contrast. Localization performance is lower for setup #1 due to pathologies and imaging artifacts in the data. On disc level, we obtain the best result for $L2/L3$ with 2.63 ± 1.52 mm (setup #2), because no severe abnormalities are present in this region in the data. Highest error rates are observed for setup #1. Due to pathologies like herniation and lumbar hyperlordosis, we reach a lower accuracy for $L5/S1$ (9.53 ± 3.67 mm) as well as for $T11/T12$ (5.64 ± 3.24 mm), where magnetic field inhomogeneities are present. Figure 2.7 (b) depicts such a case.

2.4 Discussion and Conclusion

In this chapter, we presented a novel, learning-based pipeline for semi-automatic labeling of lumbar MR scans in 3D. The main contribution lies in the generality of our method: We

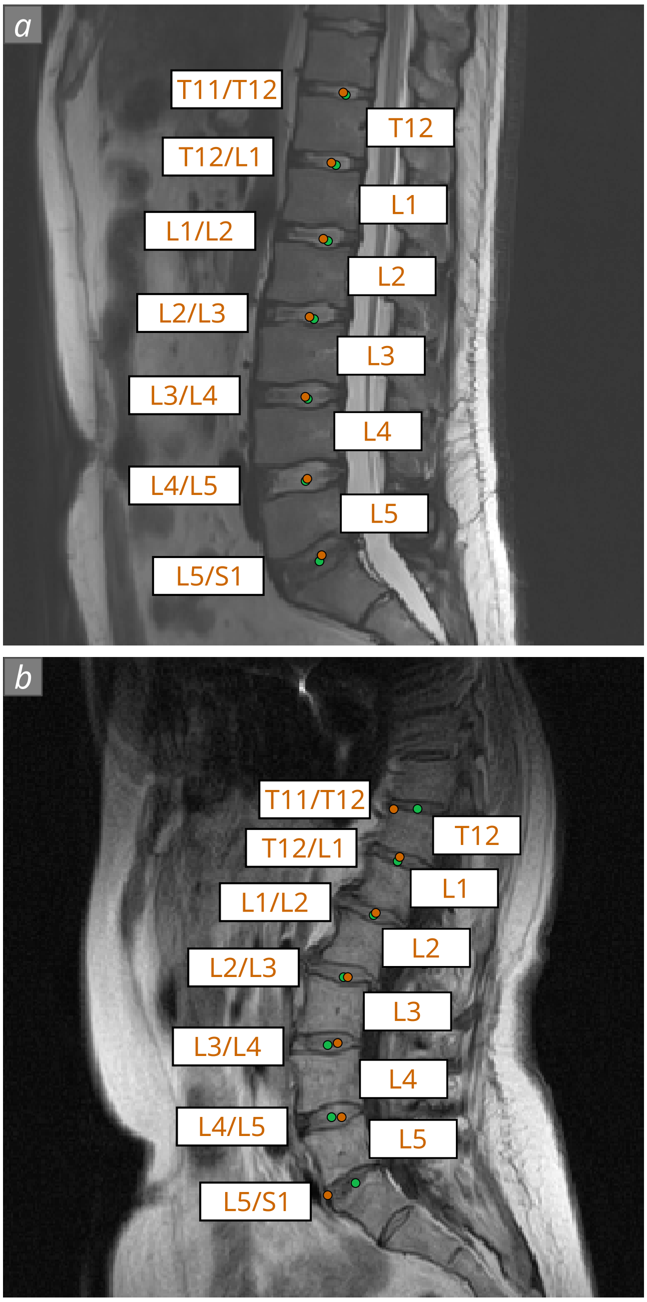


Figure 2.7: Labeling results shown on two mid-sagittal images: Our detected intervertebral disc positions (orange) and corresponding ground truth centers (green): (a) successful labeling of a healthy patient ($\mathcal{D}_3$), (b) higher position errors for $L5/S1$ and $T11/T12$ in a pathological scan with imaging artifacts ($\mathcal{D}_1$).

	p_2 [%]	p_4 [%]	p_6 [%]	p_{10} [%]	$\bar{e} \pm sd$ $[mm]$
#1	**14.29**	**40.00**	**60.71**	**84.39**	**5.78 ± 3.76**
$\mathcal{D}_1$	13.88	37.96	57.14	77.76	6.27 ± 4.18
$\mathcal{D}_2$	14.69	42.04	64.29	91.02	5.29 ± 3.22
#2	**22.51**	**64.53**	**84.99**	**97.64**	**3.82 ± 2.47**
$\mathcal{D}_3$	27.76	64.49	80.14	94.97	4.04 ± 3.14
$\mathcal{D}_4$	19.21	64.56	88.04	99.32	3.68 ± 1.94
#3	**21.32**	**59.00**	**78.46**	**92.06**	**4.45 ± 3.44**
$\mathcal{D}_1$	6.31	40.45	60.11	80.52	6.24 ± 4.12
$\mathcal{D}_2$	11.84	43.06	71.63	94.49	4.93 ± 2.67
$\mathcal{D}_4$	32.14	74.15	89.71	96.34	3.44 ± 2.99

Table 2.2: Overall performance measures per evaluated setup (bold) and corresponding dataset-specific results.

can process various imaging protocols and apply our approach also to unseen protocols, which are not covered by the training set. Furthermore, our method is significantly faster to train than recent deep learning approaches [28]. We successfully localize 84.99 % intervertebral disc centers within 6 mm and 97.64 % within 10 mm to the ground truth center, which is competitive to localization measures of state-of-the-art methods. Zukić et al. [272] report a false negative rate of 7.1 % for automatic vertebrae detection. This method is closest to ours in terms of the variability of MR data as they included T1w, T2w, and TIRM MR scans in their evaluation. Chen et al. [35] reach a position error of only 1.3 ± 0.6 mm for all intervertebral discs on Dixon data, whereby all image channels are used for training and testing. We report 3.68 ± 1.94 mm (setup #2) and 3.44 ± 2.99 mm (setup #3), whereby we did not include Dixon data in the training. On an extended set of the challenge dataset $\mathcal{D}_3$ [36], they achieve position errors between 1.8 ± 1.1 mm and 2.8 ± 6.5 mm for different cross-validation setups. We obtain disc center positions with a mean distance of 4.04 ± 3.14 mm to the expert-annotated ground truth position ($p_{10} = 94.97\%$). Overall, a higher deviation is present for $L5/S1$ and $T11/T12$, which we believe to decrease with a more enhanced refinement method. In summary, we report higher position errors with our general pipeline and evaluation setup than related works. In future work, the accuracy of disc center positions can be increased, e.g., with an improved refinement strategy and training of three-disc models on a larger, more diverse dataset. One limitation of our presented framework is its semi-automatic nature. In a follow-up work [248], we extend our method to a fully automatic system and evaluate it on a larger field of view of the spine. We introduce this approach in the next chapter. Finally, we refer to Section 5.2 for a discussion of the presented method in the context of more recent related work.

CHAPTER 3

Towards Fully Automatic Labeling of the Spine in 3D Imaging Data

This chapter is based on the following publication:

> Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. "Fully automatic cross-modality localization and labeling of vertebral bodies and intervertebral discs in 3D spinal images" in *International Journal of Computer Assisted Radiology and Surgery*, 13(10), pp. 1591–1603, 2018. DOI: 10.1007/s11548-018-1818-3

3.1 Introduction

With the growing number of MR and CT scans acquired to analyze the spine, the demand for semi- and fully-automated solutions that accelerate the time-consuming, manual spine labeling task has increased. As reviewed in Chapter 2, a common practice is to learn a dedicated method for every modality or image contrast. However, the algorithms are oftentimes not applicable to images from a different vendor or if being acquired with changed MR scan parameters [184]. Therefore, the algorithms would require retraining – a process not feasible in clinical practice and for radiology software providers. In addition, often only a few samples per imaging sequence are available to train a learning-based algorithm. This can lead to overfitting and a reduced generalization capability.

Our main vision in this work is to further develop our method presented in Chapter 2 towards a cross-modality spine localization and labeling solution that is applicable in daily clinical practice. As in the previous chapter, we aim for an algorithm with good generalization capability that does not require retraining of learned models, but can deal with the variability of scans with just a single model. More than that, our algorithm should be applicable to imaging sequences not covered by training. Our fully automatic spine labeling pipeline should be able to deal with a highly heterogeneous set of MR and CT scans, including:

- various MR *image contrasts*, including T1w, T2w, Proton Density weighted (PDw) scans,
- various MR *sequence types*, including Spine Echo, Fast Spin Echo (FSE), SPACE, TIRM, Dixon,
- and scans from *different modalities*, including MRI and CT.

We address thesis goals **G.1** and **G.2**, as in Chapter 2, and also goal **G.3** due to the inclusion of *multi-modal* data.

3.1.1 Related Work

The following literature review addresses related papers published prior to the work by Wimmer et al. [248], on which this chapter is based. We refer to Section 5.2 for a discussion of more recent works in the field of spine labeling.

Most fully and semi-automatic disc and vertebra localization and semantic labeling methods are trained on a specific modality, image sequence, or parameter setting [13, 184], as briefly reviewed in Section 2.1. Many CT-specific works have been presented during the past years for both healthy and highly pathological scans [38, 69, 70, 149, 221]. For MR data, contrast and sequence specific methods are proposed, for example on T1w [156, 199] or T2w data [12, 36, 49, 141] or methods that fuse various contrast information [167]. Protocol-specific algorithms, for example on Dixon data using one image channel [79] or fusing multiple channels [35, 98, 131], also exist. Most algorithms require retraining to be applicable to data acquired with different parameters or other modalities, as shown by Kelm et al. [156]. Another method requires modality-specific parameters to deal with CT and T2w MR data [42].

Algorithms dealing with a wider range of imaging contrasts and modalities have also been presented. Štern et al. [220] facilitate edge and gradient information for fully automated spinal centerline detection in T1w, T2w MR, and CT scans. The centerline refers to the "*curve in 3D that passes through the centre of each vertebral body*", according to Štern et al. [220]. Intensity profile analysis along the centerline of the spine provides disc and vertebrae center positions, but an automated assignment of corresponding anatomical labels is not performed. Zhan et al. [262] employ a hierarchical detection strategy and model the spatial relation between discs and vertebrae with articulated models on CT and T2w MR scout scans. A semi-automatic localization and labeling approach [83] recently applied to T1w and T2w MR builds a subject-specific density model from a click position in a predefined vertebra and performs distribution-matching for candidate detection without external training. Wang et al. [241] address disc and vertebrae segmentation as a boundary regression problem for CT and T1w and T2w MR scans. However, anatomical labeling of segmented vertebrae has not been performed. Recently, algorithms employing deep learning in the medical field [136] and also specifically for spine labeling have been presented. Fully Convolutional Networks (FCNs) for disc localization and segmentation on T2w MR [37] and neural networks combined with graphical models on 2D T1w and T2w MR scans are introduced [60]. Cai et al. [27, 28] go in the direction of a multi-modal solution for T1w, T2w MR, and CT scans using restricted Boltzmann machines and feature fusion. However, the imaging modality is deduced from Digital Imaging and Communications in Medicine (DICOM) tags for labeling an unseen scan.

In many MR sequences the ribs cannot be reliably detected due to low image contrast or a restricted field of view focusing on the spine only. To unambiguously distinguish discs and vertebrae, all reviewed MR labeling algorithms rely on the presence of an anchor point. In CT data this is typically achieved by searching for vertebrae connected to ribs, the sacrum and/or cervical vertebrae [149].

3.1.2 Scope and Contribution

The method presented in this chapter significantly extends the previous work [247] summarized in Chapter 2. There we addressed thesis goals **G.1** and **G.2**, and presented a semi-automatic spine labeling approach on T1w, T2w, and Dixon MR scans. The suggested algorithm requires an initial click position and its corresponding anatomical label as user input. To overcome this limitation, we propose:

- An entropy-based sacrum model and a CNN for the task of fully automatic localization of the sacral region in sagittally reconstructed 3D scans. The algorithm works on various MR image contrasts and sequences *without retraining* the algorithm, including completely unseen image contrasts and sequences, as well as CT scans not covered by training. Thus, we additionally address thesis goal **G.3** in this chapter. For successful labeling, we require the sacrum to be on a scan as reliable anchor position.

- Subsequent, fully automatic labeling of vertebrae and intervertebral discs in this highly heterogeneous set of spine scans. We extend our three-disc model labeling approach [247] presented in Chapter 2 by an improved position-refinement method that utilizes a *3D vertebral canal template*.

- Extensive evaluation of our pipeline on highly heterogeneous *public datasets* showing the lumbar and thoracolumbar spine. Compared to our previous work [247] introduced in Chapter 2, our algorithm is evaluated on more than 100 additional scans, including several new MR image contrasts and CT scans.

3.2 Methods

To achieve a cross-modality solution, we again employ ETMs, as introduced by Zambal et al. [260] (see Section 2.2.1 for details). We build *local models* covering vertebrae sequences that can capture local intensity variations within scans, e.g., due to pathologies or acquisition from different scanners. We employ ETMs for seed point localization and iterative labeling, which enables using a single learning-based pipeline without parameterizing it to different imaging modalities.

Our proposed pipeline comprises three main parts: (a) *seed point localization*, (b) *labeling*, and (c) *refinement*. First, we introduce an ETM-based sacral region classifier for seed point localization in Section 3.2.1, which we require to initialize the labeling. Next, we recapture local three-disc models for iterative labeling in Section 3.2.2. To increase the center position accuracy of localized positions, we introduce a vertebral canal template for position refinement in Section 3.2.3. The combination of the three parts to the complete, fully automatic labeling pipeline is presented in Section 3.3. Figure 3.1 gives an overview of the system.

Notation: We use the same mathematical notation as introduced in the previous chapter (see Section 2.2). We denote a volume $V \subset \mathbb{R}^3$ as the 3D reconstruction of a stack of 2D sagittal slices and $\boldsymbol{p} = (x, y, z)$ as a position within V. The labelset $\Lambda = (S1, L5/S1, L5, \ldots, C3, C2/C3)$ denotes the ordered set of anatomical labels of vertebrae and intervertebral discs in the human spine, from the sacrum $S1$ to the topmost cervical disc $C2/C3$. Consequently, a position labeled with a certain anatomical label is given by $\boldsymbol{p}^{\lambda_l} \in V$, with $\lambda_l \in \Lambda, l = 1, \ldots, |\Lambda|$. Again, we denote a model-matched position with $\tilde{\boldsymbol{p}}$, a refined position with $\hat{\boldsymbol{p}}$, and the annotated ground truth center of an intervertebral disc or vertebra in a scan with $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$. Finally, we define a classifier Φ, which classifies 2D patches into “sacrum” / “non-sacrum”.

3.2.1 ETM-Based Sacral Region Classifier for Seed Point Localization

To overcome the problem of seed point localization in highly heterogeneous MR and CT data, we propose the following steps: First, we build a 3D ETM M^S covering the sacral region, which can normalize the data to the reduced number of t target gray levels. Then, we introduce a modality-independent sacral region classifier Φ that classifies a normalized 2D patch Q into “sacrum” / “non-sacrum”, i.e., a patch showing the sacrum or another region of a scan. The output of this stage is the learned classifier Φ that maps a patch Q to a classification score, i.e., $\Phi(Q) : \mathbb{R}^2 \mapsto [0, 1]$. $\Phi(Q)$ refers to the probability of Q being a sacrum patch.

3D Sacrum ETM $\boldsymbol{M^S}$. We build a modality-independent local 3D sacrum entropy model M^S from sparse landmarks covering the lower lumbar and sacral region $L4/L5$ to $S1$. Figure 3.2 (a) shows the extracted landmarks, projected onto the mid-sagittal slice. M^S is built around the four annotated tissue centers $\boldsymbol{d}^{\boldsymbol{\lambda_{S1}}}$, $\boldsymbol{d}^{\boldsymbol{\lambda_{L5/S1}}}$, $\boldsymbol{d}^{\boldsymbol{\lambda_{L5}}}$, and $\boldsymbol{d}^{\boldsymbol{\lambda_{L4/L5}}}$.

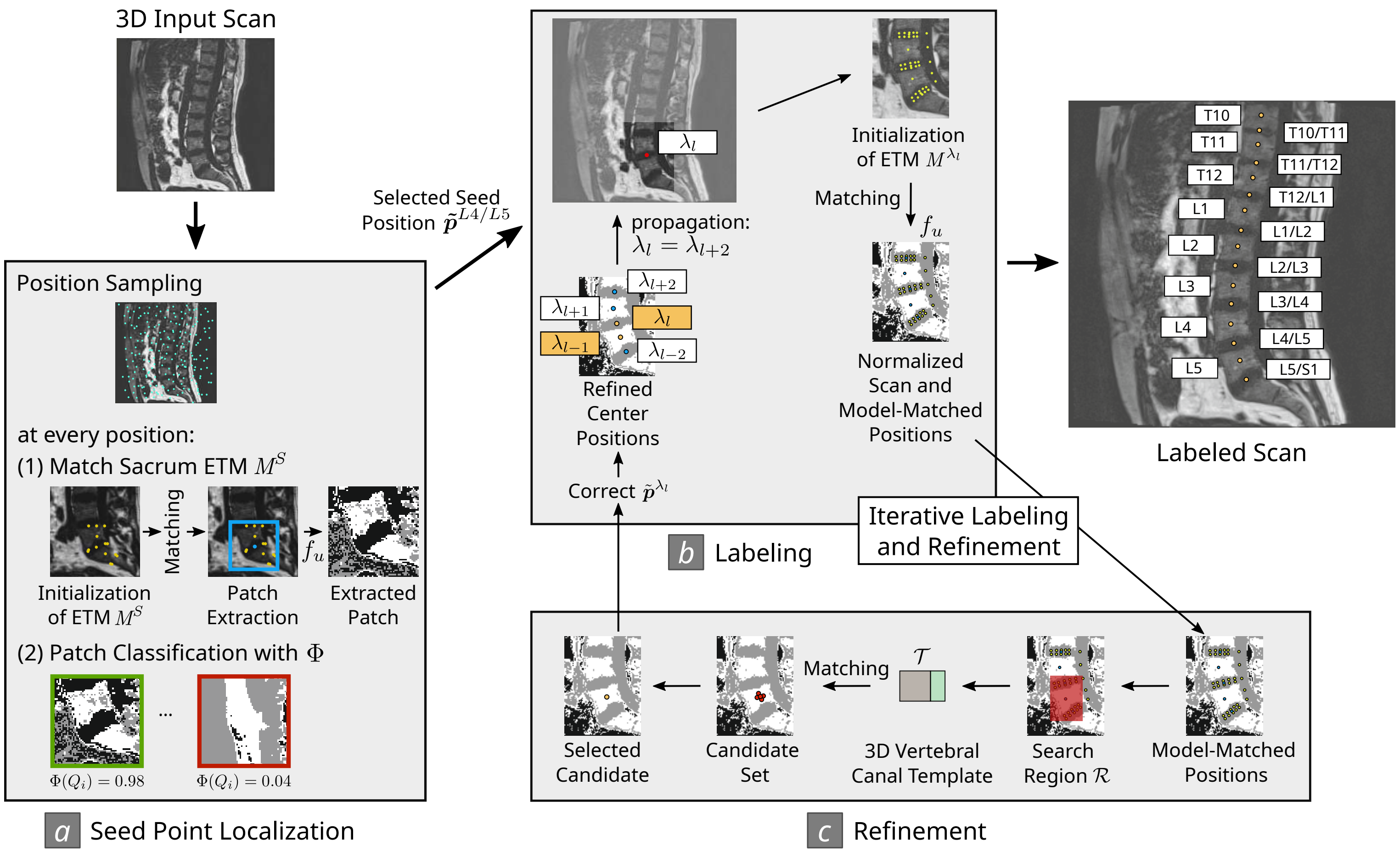


Figure 3.1: Schematic illustration of fully automatic spine labeling of a 3D scan V with our proposed pipeline.

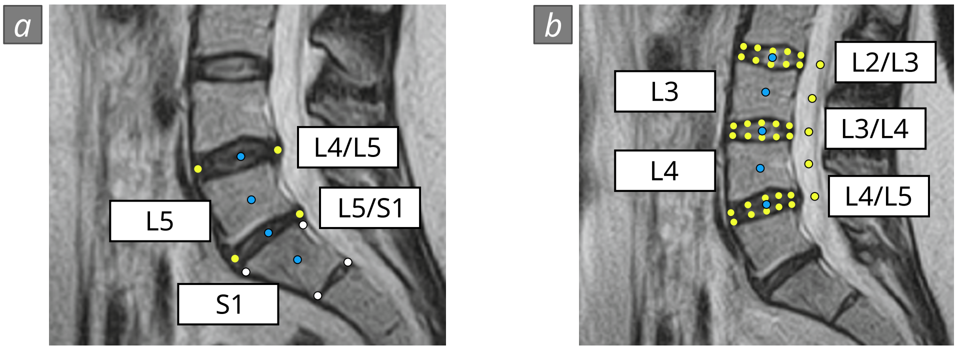


Figure 3.2: Schematic illustration of landmarks extracted for building (a) sacrum model M^S, and (b) sample three-disc model M^{λ_l}, $\lambda_l = L3/L4$ for labeling. Intervertebral disc and vertebrae center positions are depicted in blue, sacrum corner positions in white, remaining model landmark positions in yellow.

Further, four landmarks at the corners of the $S1$ vertebra as well as positions at the left and right boundaries of discs are extracted from the topmost and bottommost slices in the sagitally reconstructed volume V where the sacrum is visible. In Figure 3.2 (a), tissue centers are shown in blue, sacrum corner positions in white, and the remaining landmarks corresponding to discs in yellow. From this set of landmarks, we train the sacrum ETM M^S following Section 2.2.1.

Sacral Region Classifier Φ. Next, we employ the sacrum ETM M^S to normalize 2D patches used as input to train a modality-independent sacral region classifier Φ. To train such a classifier, we require a set of positive patches, i.e., patches that show the sacrum, and a set of negative ones, which show other regions in a scan. We will denote a positive and negative patch with Q^+ and Q^-, respectively.

We extract patches in the following way: For every volume $V_k \in \mathcal{D}_{tr}$, we sample n^+ positions $\boldsymbol{p}_i^+$, $i = 1, \ldots, n^+$ that lie within a certain distance d^+ of the annotated sacrum center $\boldsymbol{d}^{\boldsymbol{\lambda}_{S1}}$, formally: $|\boldsymbol{p}_i^+ - \boldsymbol{d}^{\boldsymbol{\lambda}_{S1}}| \leq d^+$, and n^- positions $\boldsymbol{p}_j^-$, $j = 1, \ldots, n^-$ anywhere in the volume with a minimum distance d^- to the ground truth position $\boldsymbol{d}^{\boldsymbol{\lambda}_{S1}}$, formally: $|\boldsymbol{p}_j^- - \boldsymbol{d}^{\boldsymbol{\lambda}_{S1}}| \geq d^-$. We place an instance of the learned sacrum model M^S at every position $\boldsymbol{p} \in \{\boldsymbol{p}_i^+, \boldsymbol{p}_j^-\}$ and match M^S locally. The matching is performed as described earlier in Section 2.2.1. Then, we extract positive and negative 2D patches Q_i^+ and Q_j^- of size $m_s \times m_s$ at the model-matched position $\tilde{\boldsymbol{p}}^{L5/S1}$ and apply the corresponding obtained intensity mapping f_u to t target values. We choose 2D over 3D patches to keep the complexity and computational cost of the classifier low. Samples of normalized patches are given in Figure 3.3.

We utilize a CNN for patch classification on the extracted 2D ETM patches and denote the trained model with Φ_E^C. Φ_E^C follows a basic LeNet-5 [119] architecture with different numbers of convolutional layers, followed by pooling and dropout layers to prevent the

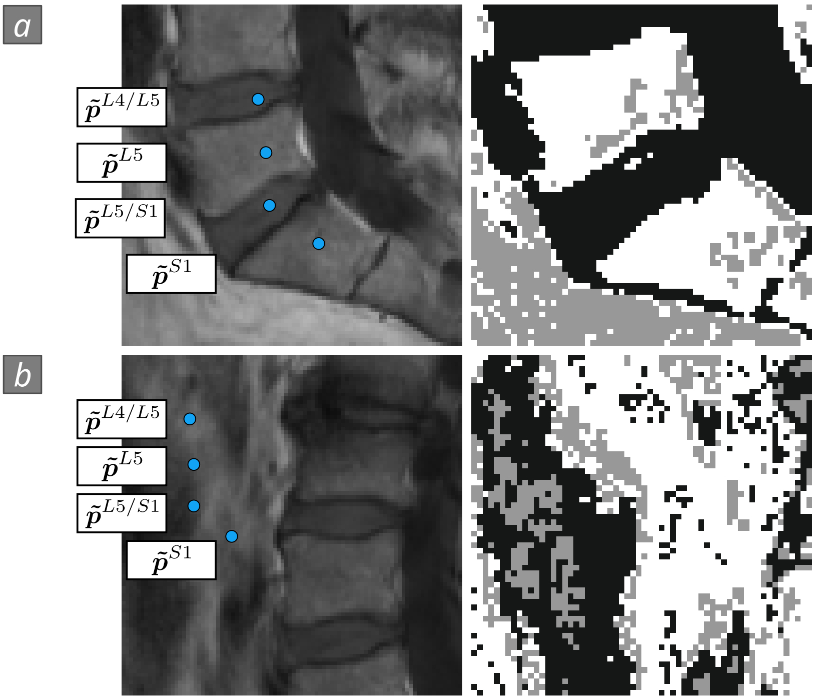

a
$\tilde{\boldsymbol{p}}^{L4/L5}$
$\tilde{\boldsymbol{p}}^{L5}$
$\tilde{\boldsymbol{p}}^{L5/S1}$
$\tilde{\boldsymbol{p}}^{S1}$
b
$\tilde{\boldsymbol{p}}^{L4/L5}$
$\tilde{\boldsymbol{p}}^{L5}$
$\tilde{\boldsymbol{p}}^{L5/S1}$
$\tilde{\boldsymbol{p}}^{S1}$

Figure 3.3: Sagittal slices from raw data showing model-matched landmark positions $\tilde{\boldsymbol{p}}^{S1}$, $\tilde{\boldsymbol{p}}^{L5/S1}$, $\tilde{\boldsymbol{p}}^{L5}$, $\tilde{\boldsymbol{p}}^{L4/L5}$ in blue (left) and corresponding extracted patch at $\tilde{\boldsymbol{p}}^{L5/S1}$ from the normalized scan V' (right): (a) Sample of a positive patch Q_i^+ where M^S is matched in the sacral region. (b) Matching M^S at a position outside of the sacral region results in a negative sample Q_j^-.

model from overfitting [217]. Apart from Φ_E^C, we train other classifiers with different patch normalization schemes for comparison, following the same patch extraction scheme as described in this section. CNN parameter search and training details for the other classifiers are explained in detail in Section 3.4.2.

3.2.2 Three-Disc Models M^{λ_l} for Labeling

For modality-independent labeling of the spine, we build local three-disc ETMs M^{λ_l}, from sparse landmarks around a middle disc with label λ_l, including its adjacent upper and lower discs. We follow the same scheme for building the models as presented in the previous chapter (see Section 2.2.2 and Figure 2.3). Figure 3.2 (b) gives an example of the landmarks extracted for the $L3/L4$ model. Again, we apply the three-disc models in an iterative labeling pipeline where we match one model after the other to the spine, as described in Section 3.3.

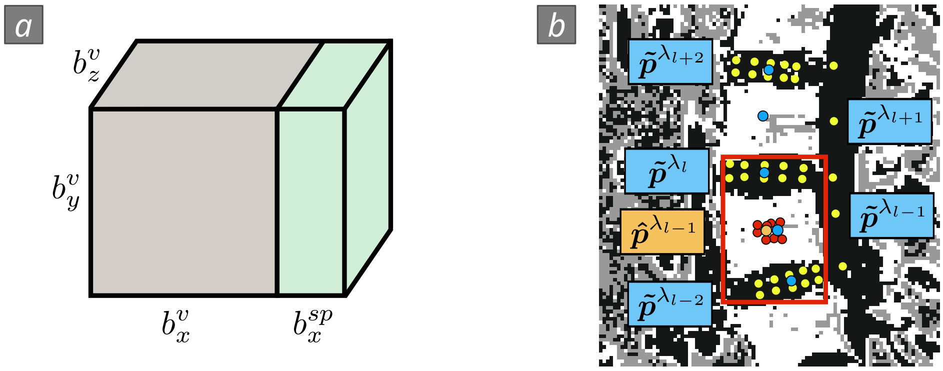

a
b_z^v
b_y^v
b_x^v
b_x^{sp}
b
$\tilde{\boldsymbol{p}}^{\lambda_{l+2}}$
$\tilde{\boldsymbol{p}}^{\lambda_{l+1}}$
$\tilde{\boldsymbol{p}}^{\lambda_l}$
$\hat{\boldsymbol{p}}^{\lambda_{l-1}}$
$\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$
$\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$

Figure 3.4: Template matching: (a) Schematic illustration of template $\mathcal{T}$ for refinement, comprising the vertebral body (gray) and spinal canal region (green). (b) Positions involved in the matching and refinement: model-matched positions of tissue centers $\tilde{\boldsymbol{p}}^{\lambda_{l-2}}$, $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$, $\tilde{\boldsymbol{p}}^{\lambda_l}$, $\tilde{\boldsymbol{p}}^{\lambda_{l+1}}$, and $\tilde{\boldsymbol{p}}^{\lambda_{l+2}}$ (blue), remaining model-matched positions $\tilde{\boldsymbol{p}}$ (yellow), candidates for refinement (red), and refined vertebra center $\hat{\boldsymbol{p}}^{\lambda_{l-1}}$ (orange). The search region $\mathcal{R}$ is illustrated by the red rectangle.

3.2.3 Vertebral Canal Template for Vertebra Position Refinement

We propose an intensity-based template matching approach based purely on the normalized data V' to account for tissue center position errors due to model mismatches of the sparse three-disc ETMs M^{λ_l}. To increase the robustness of the refinement, we apply our method to the model-matched lower vertebra position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$ in this chapter instead of the middle disc position $\tilde{\boldsymbol{p}}^{\lambda_l}$, as in Chapter 2.

3D Vertebral Canal Template $\mathcal{T}$. First, we construct a template $\mathcal{T}$ for the refinement of the lower vertebra with label λ_{l-1}. It comprises two neighboring box-shaped regions, where one approximates the size of the vertebral body of vertebra λ_{l-1} and the other one the spinal canal adjacent to the vertebral body (see Figure 3.4 (a)). The sizes of both regions depend on vertebra λ_{l-1} and are based on spine morphometry measures combined from several studies [67, 171, 172]. We denote the size of the vertebral body region in $\mathcal{T}$ with $b_x^v \times b_y^v \times b_z^v$. It ranges from 34 × 25 × 34 mm for $L5$ to 22 × 17 × 22 mm for $T2$. Corresponding canal widths are 15 mm and 12 mm for $L5$ and $T2$, respectively, which we denote with b_x^{sp}. Both template regions are initialized with intensity values corresponding to the t target gray levels in the normalized volume V'. We determine the *vertebra intensity mode*, i.e., the most frequent gray value, in a local 3 × 3 × 3 window at the model-matched lower vertebra position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$ obtained after matching M^{λ_l} to the spine and assign it to the vertebra region in $\mathcal{T}$. All remaining target gray levels are considered for the canal region. In this way, we build the template in a “one-versus-rest“ manner, as the intensity mapping in a vertebra is rather homogeneous compared to the spinal canal, which can exhibit different intensities due to the presence of the spinal cord.

Template matching. In search region $\mathcal{R}$ (see Figure 3.4 (b), red rectangle), constrained by the convex hull of the matched three-disc model M^{λ_l}, we place $\mathcal{T}$at every voxel position $\boldsymbol{p}_i$ in $\mathcal{R}$. We calculate the *template overlap* $\Theta_i^{\mathcal{T}}$ with the current underlying normalized volume V' at every voxel $\boldsymbol{p}_i$:

$$\Theta_i^{\mathcal{T}} = \frac{\Theta_i^v}{\Theta^v} + \frac{\Theta_i^{sp}}{\Theta^{sp}} \tag{3.1}$$

Θ_i^v and Θ_i^{sp} are the number of overlapping intensity values of the vertebral body and spinal canal template region, respectively, with the underlying normalized volume V'. Θ^v and Θ^{sp} are the sizes in voxels of the vertebral body and the spinal canal region, respectively. The best candidate $\boldsymbol{p}^*$ in $\mathcal{R}$ is selected as the one having maximal overlap $\Theta_i^{\mathcal{T}}$:

$$\boldsymbol{p}^* = \underset{\boldsymbol{p}_i \in \mathcal{R}}{\operatorname{argmax}}(\Theta_i^{\mathcal{T}}) \tag{3.2}$$

3.3 Fully Automatic Labeling of an Unseen Scan

This section combines the models and methods introduced in Section 3.2 to the full pipeline for labeling a scan V. Figure 3.1 gives an overview of the complete system.

3.3.1 Seed Point Localization

First, we sample positions $\boldsymbol{p}_i = (x_i, y_i, z_i)$, $i = 1, \ldots, n$ uniformly in V, as shown in Figure 3.1 (a). We initialize an instance of the learned sacrum ETM M^S at every position $\boldsymbol{p}_i$, match M^S locally to the unseen scan, and extract a patch Q_i at the model-matched position $\tilde{\boldsymbol{p}}_i^{L5/S1}$. Next, we apply the obtained intensity mapping f_u and classify Q_i into "sacrum" / "non-sacrum" with classifier Φ. We consider Q_i a sacrum patch if $\Phi(Q_i) \geq 0.5$ and select position $\tilde{\boldsymbol{p}}_i$ corresponding to the patch with highest probability as candidate:

$$\tilde{\boldsymbol{p}}_i = \underset{i \in \{1,\ldots,n\}}{\operatorname{argmax}} \{\Phi(Q_i)\} \tag{3.3}$$

Finally, we select the corresponding position $\tilde{\boldsymbol{p}}^{L4/L5}$ that indicates the center of disc $L4/L5$ from the matched sacrum model and use it as seed point for the subsequent iterative labeling and refinement steps.

3.3.2 Iterative Labeling and Refinement

The obtained $L4/L5$ position is used as anchor for anatomical labeling of the rest of the spine. This includes detecting vertebrae and disc center positions, as well as assigning their corresponding anatomical labels λ_l.

As in Chapter 2, we iteratively match three-disc models M^{λ_l} to the spine. For position refinement, we apply the proposed vertebrae template matching. We start iterative model matching at $\tilde{\boldsymbol{p}}^{\lambda_l}$ and repeat the following steps:

1. Initialize the three-disc model M^{λ_l} for current disc λ_l: We place M^{λ_l} at the position obtained from the previous model-matching step and match it (see Figure 3.1 (b)). Next, we apply the obtained intensity mapping f_u which results in the normalized scan. Further, we retrieve the set of model-matched labeled positions $\{\tilde{\boldsymbol{p}}^{\lambda_{l-2}}, \tilde{\boldsymbol{p}}^{\lambda_{l-1}}, \tilde{\boldsymbol{p}}^{\lambda_l}, \tilde{\boldsymbol{p}}^{\lambda_{l+1}}, \tilde{\boldsymbol{p}}^{\lambda_{l+2}}\}$ (see positions in Figure 3.4 (b) in blue color).

2. *Refine* the lower vertebra position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$ as given by the model matching according to the approach presented in Section 3.2.3: We build the 3D vertebral canal template $\mathcal{T}$ with the size depending on the current vertebra λ_{l-1}. Next, we match $\mathcal{T}$ within the search region $\mathcal{R}$. To increase the robustness of the refinement, we select the set of n positions $\{\boldsymbol{p}_i^*\}$, $i = 1, \ldots, n$ with the n highest template overlaps according to Equation 3.2 as the candidate set.

3. Select the best candidate from the set: We assume that the positions obtained from model matching are reliable, i.e., that the model-matched position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$ lies inside the corresponding vertebra. Hence, the final refined position $\hat{\boldsymbol{p}}^{\lambda_{l-1}}$ should lie in proximity to the model-matched position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$. We check for every candidate position $\boldsymbol{p}_i^*$, $i = 1, \ldots, n$, starting with the best candidate, i.e., the one with the highest template overlap, the condition $| \tilde{\boldsymbol{p}}^{\lambda_{l-1}} - \boldsymbol{p}_i^* | \leq \mu$; that is, we check whether the candidate position lies within distance μ. The distance μ is based on spine morphometry measures [67, 171, 172] and varies between 25 mm for $L5$ to 16 mm for $T2$. The first candidate meeting this criterion is selected as refined lower vertebra position $\hat{\boldsymbol{p}}^{\lambda_{l-1}}$. In case none of the candidates $\{\boldsymbol{p}_i^*\}$ fulfills the criterion, we set $\hat{\boldsymbol{p}}^{\lambda_{l-1}} := \tilde{\boldsymbol{p}}^{\lambda_{l-1}}$, i.e., we select the model-matched position as the refined vertebra center.

4. Correct $\tilde{\boldsymbol{p}}^{\lambda_l}$: To reduce landmark position errors that may have occurred in the z-axis of the sagittally reconstructed scan after matching M^{λ_l}, we correct the model-matched position of the middle disc $\tilde{\boldsymbol{p}}^{\lambda_l}$. We calculate the distance between the refined lower vertebra position $\hat{\boldsymbol{p}}^{\lambda_{l-1}}$ and the corresponding model-matched position $\tilde{\boldsymbol{p}}^{\lambda_{l-1}}$ in z and correct $\tilde{\boldsymbol{p}}^{\lambda_l}$ in z by that distance. The obtained position is the refined middle disc position $\hat{\boldsymbol{p}}^{\lambda_l}$.

5. Propagation: Place the next three-disc model at the next unlabeled disc position λ_{l+2}: $\lambda_l = \lambda_{l+2}$, and start from (1).

We stop if no more discs are found. This is the case if either the border of a scan or the topmost disc $C2/C3$ is detected. The final result of the algorithm is the set of labeled, refined positions $\{\hat{\boldsymbol{p}}^{\lambda_i}\}, i = 1, \ldots, K$ where K corresponds to the number of tissues in V.

3.4 Experimental Setup

3.4.1 Datasets

We evaluate our pipeline on eight different datasets $\mathcal{D}_i$, $i = 1, \ldots, 8$, comprising 184 sagittally reconstructed scans V with 1659 annotated vertebra and disc centers. Table 3.1 gives a detailed overview of the data characteristics. The datasets represent various MR sequences and image contrasts that had been acquired on different scanners exhibiting highly anisotropic voxel sizes, as well as CT scans. $\mathcal{D}_1$ and $\mathcal{D}_2$ are private collections. The remaining datasets had been released with recent works [28, 35, 36, 42, 272] and/or made publicly available through the SpineWeb [215] initiative.

Pathologies: The following pathologies are present in at least half of the scans in $\mathcal{D}_1$ and $\mathcal{D}_2$: fractures, disc herniation, scoliosis, lordosis, and combinations of those. $\mathcal{D}_8$ [272] includes two scans without pathologies and 14 scans with pathologies: stenosis (one scan), scoliosis (one scan), spondylolisthesis (two scans), vertebral fracture (one scan), vertebral fracture and spondylolisthesis (one scan), other pathologies not diagnosable from the provided vertebra segmentation (seven scans), and vertebral fracture and other pathologies (one scan). The multi-modality datasets $\mathcal{D}_6$ and $\mathcal{D}_7$ by Cai et al. [28] comprise healthy cases and patients with minor spondylosis/fractures. No details about present pathologies are provided for datasets $\mathcal{D}_3$, $\mathcal{D}_4$, and $\mathcal{D}_5$

3.4.2 Evaluation Setup

Our evaluation setup demonstrates the applicability of our method to *unseen sequences and contrasts* that are not covered by the training set $\mathcal{D}_{tr}$. We conduct the training of all models on set $\mathcal{D}_{tr} = \mathcal{D}_1 \cup \mathcal{D}_2$ where only T1w and T2w MR scans are present. The full labeling pipeline is evaluated on $\mathcal{D}_{test} = \mathcal{D}_3 \cup \mathcal{D}_4 \cup \mathcal{D}_5 \cup \mathcal{D}_6 \cup \mathcal{D}_7 \cup \mathcal{D}_8$. This set comprises scans at various spatial resolutions and unseen sequences and image contrasts.

To further validate the generalization capability of our approach, we provide additional results when training on the set of CT scans $\mathcal{D}_6$, and evaluating the full pipeline on the remaining MR datasets $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$, $\mathcal{D}_5$, $\mathcal{D}_7$, and $\mathcal{D}_8$. We summarize and discuss these results in Section 3.5.4.

ETM Training

This section describes ETM parameter selection [260] for the sacrum model M^S and three-disc models M^{λ_l}. For the target levels t, we focus on $t \in \{2, 3\}$ target levels, because we aim for homogeneous intensity mappings for discs, vertebrae, and the spinal cord. With only $t = 2$ target values, we encounter the problem that the intensity reduction is too strong in regions exhibiting low contrast, e.g., at the border of the volume, or with scans in a high intensity range. This results in a degeneration of mappings and hence a loss of information about the underlying anatomy. This is not the case for $t = 3$, where desirable mappings for vertebrae and discs are obtained. For the source levels r, we vary

ID	Modality & Contrast/Sequence	In-plane Resolution xy	Slices per Scan	Pixel size xy $[mm^2]$	Slice Thickness $[mm]$	No. Scans (No. Annotations)	Description
$\mathcal{D}_1$	MR T1w TSE	324 × 324 - 512 × 1436	11 - 16	0.586 - 1.173	4.4 - 6	13 (291)	Our T1w Dataset: scans from 12 different patients, including 3 full spine scans, acquired on Philipps Achieva and Panorama scanners. Annotation: disc and vertebrae centers, $C2/C3$ - $L5/S1$ (by medical expert)
$\mathcal{D}_2$	MR T2w TSE	320 × 320 - 512 × 1436	13 - 14	0.586 - 1.187	4.4 - 6	10 (244)	Our T2w Dataset: scans from 10 different patients, including 3 full spine scans, acquired on Philipps Achieva and Panorama scanners. Annotation: disc and vertebrae centers, $C2/C3$ - $L5/S1$ (by medical expert)
$\mathcal{D}_3$	MR T2w	304 × 304, 305 × 305	48, 39	1.25	2.0	15 (105)	MICCAI Challenge 2015 Training Dataset [36]: 15 different patients, acquired on one scanner. Annotation: disc centers, $T11/T12$ - $L5/S1$
$\mathcal{D}_4$	MR T2w TSE	305 × 305	39	1.25	2.0	23 (161)	Chu Dataset [42]: public data, acquired on 1.5 T Siemens scanner. Annotation: vertebrae centers, $T11$ - $L5$
$\mathcal{D}_5$	MR Dixon	256 × 256	36	1.25	2.0	32 (280)	Dixon Dataset [35]: 8 different patients, whereby 4 image channels are acquired per patient: fat saturated, water saturated, inn-phase and opposed-phase. Annotation: disc centers, $T11/T12$ - $L5/S1$
$\mathcal{D}_6$	CT	512 × 512	28 - 86	0.25 - 0.408	1.5 - 3.0	19 (63)	Cai Dataset [28]: 20 different patients, whereby only sagittally reconstructed CT images are used in this work. One scan shows the cervical spine, but is kept for seed point localization evaluation. Annotation (lumbar): 3-6 vertebrae centers, $L4$ - $S1$ resp. $L1$ - $S1$
$\mathcal{D}_7$	MR T1w TSE, T1w FLAIR, T2w CUBE, PDw TSE	320 × 320 - 640 × 640	12 - 160	0.406 - 0.859	0.5 - 5.0	56 (355)	Cai Dataset [28]: 20 different patients, whereby only sagittal MR scans are considered in this work; four scans show the cervical spine, but are kept for seed point localization evaluation. Annotation (lumbar): 6-9 vertebrae centers, $L1$ - $S1$ resp. $T10$ - $S1$
$\mathcal{D}_8$	MR T1w FSE, T1w TSE, T2w FSE, T2w TSE, TIRM	320 × 320 - 768 × 768	12 - 31	0.47 - 1.188	3.0 - 4.4	16 (160)	Zukić Dataset [272]: public dataset, acquired on several scanners in different hospitals, whereby only the 16 sagittal scans are included in this work (one coronal scan excluded). Annotation: 7-17 vertebrae centers per scan, $T12$ - $S1$ resp. $T2$ - $S1$

Table 3.1: Detailed dataset overview.

$r \in \{70, 75, \ldots, 130\}$. For $t = 3$, the best results are obtained with $r = 110$ source levels using leave-one-out cross-validation on $\mathcal{D}_{tr}$. This combination provides a good initial quantization of intensities, where relevant intensity changes are preserved and image noise is removed. We do not obtain the same intensity mapping for one tissue in all sequences. The intensity ranges in the original scans vary, but homogeneous mappings within the tissues are obtained with the learned models M^{λ_l} and M^S.

Sacral Region Classifier Training

From every volume $V_k \in \mathcal{D}_{tr}$, $n^+ = 100$ patches Q_i^+ are sampled within $d^+ \leq 10$ mm from the annotated sacrum center and $n^- = 500$ negative patches Q_j^- with $d^- \geq 50$ mm. The patch size is empirically set to $m_s = 60$ pixels. This ensures that the $L4/L5 - S1$ region is well covered by a patch. We split $\{Q_i^+, Q_i^-\}$ into 90 % of patches belonging to the training set and the remaining 10 % to the validation set. To increase robustness, we perform data augmentation and additionally added 15 patches per class and per scan in $\mathcal{D}_{tr}$ to the validation set. We perform random cropping to a size between 50 × 50 and 70 × 70 pixels and rescaling to 60 × 60 pixels.

Training: The CNN Φ_E^C is trained on ETM patches, as proposed in Section 3.2.1. To justify our choice, we compare its performance to Support Vector Machines (SVMs) on the same data and denote this model with Φ_E^S. Further, we train two additional CNNs on raw data with different preprocessing strategies. One CNN Φ_S^C uses standard scaling and another one Φ_W^C applies whitening [113]. We optimize the parametrization of CNN classifiers Φ_E^C, Φ_S^C, and Φ_W^C via grid search by varying the following parameters:

- dropout: $\{0.1, 0.2, 0.3, 0.4, 0.5\}$
- number of consecutive convolutional layers before pooling layer: $\{1, 2, 3\}$
- size of filter in convolutional layers: $\{3, 5\}$
- number of filters in convolutional layers: $\{32, 64\}$
- number of dense layer units as percentage of its previous layer units: $\{10\%, 20\%, 25\%, 30\%, 50\%\}$

We train the CNNs by optimizing the cross-entropy loss using stochastic gradient descent with Nesterov momentum [91] of 0.9 and a learning rate lr of 0.01 for 100 epochs. Rectified linear units are used as non-linear functions, except for the last layer where we apply the softmax activation function. For the SVM Φ_E^S with Radial Basis Function (RBF) kernel, we perform grid search for parameters C and γ with $\{C, \gamma\} \in \{0.0001, 0.001, \ldots, 10000\}$. We use the same predefined training/validation split for training SVMs and CNNs and select the best classifiers based on the performance on the validation set:

- Φ_E^C: ETM data, dropout = 0.1 (input layer) and 0.5 (dense layers), two convolutional layers, filter size = 5, number of filters = 64, number of dense layer units as percentage of its previous layer units = 20 %
- Φ_E^S: RBF kernel with $\gamma = 0.0001$, $C = 10$

3.4.3 Metrics

We evaluate the seed point localization and labeling performance based on two measures defined in Section 2.3.3: the intervertebral disc localization performance p_ϵ and the disc center accuracy $\bar{e}$. However, in the previous chapter we focused only on intervertebral discs while now we are concerned with vertebrae positions as well. Hence, we reformulate the metrics more general on a *tissue level* as follows:

- p_ϵ − **Tissue localization accuracy:** p_ϵ explains how many detected positions $\hat{\boldsymbol{p}}^{\lambda_l}$ lie *within* ϵ *mm* to the respective annotated tissue center $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$, with $\epsilon \in \{2, 4, 6, 10, 15\}$.
- $\bar{e} \pm sd$ − **Tissue center accuracy:** The distance from the localized position $\hat{\boldsymbol{p}}^{\lambda_l}$ to the ground truth $\boldsymbol{d}^{\boldsymbol{\lambda_l}}$ is given by the mean position error $\bar{e}$ and standard deviation sd.

3.4.4 Implementation Details

Our framework is implemented in Python, using the Lasagne deep learning library [54]. Training and testing of CNNs is conducted on an Nvidia GeForce GTX 970 GPU.

3.5 Results and Discussion

The following section summarizes the results of fully automatic seed point detection and labeling on $\mathcal{D}_{test}$.

3.5.1 Seed Point Localization Performance

The performance of the learned classifiers on $\mathcal{D}_{test}$ is compared in Figure 3.5, where the distance from the annotated ground truth center to the localized seed is shown. From the total number of 161 scans in $\mathcal{D}_{test}$, 156 show the lower spine including the sacrum, while the remaining five are pure cervical scans. We correctly detect seeds in 146/156 scans with our proposed classifier Φ_E^C, representing a detection rate of 93.6 %. Seeds are completely missed in only two scans in $\mathcal{D}_7$, i.e., in 1.3 % of all scans. We detect a few outliers, which are summarized in Table 3.2 and visualized in Figure 3.6. No sacrum seeds are detected in the five cervical scans in $\mathcal{D}_6$ and $\mathcal{D}_7$, which reflects the desired behavior. Hence, no false positive predictions are made. Only a few shifts occur where $S1/S2$ or $L4/L5$ are confused with $L5/S1$ ($\mathcal{D}_3$, $\mathcal{D}_4$, $\mathcal{D}_7$, $\mathcal{D}_8$) or due to a pathological scan

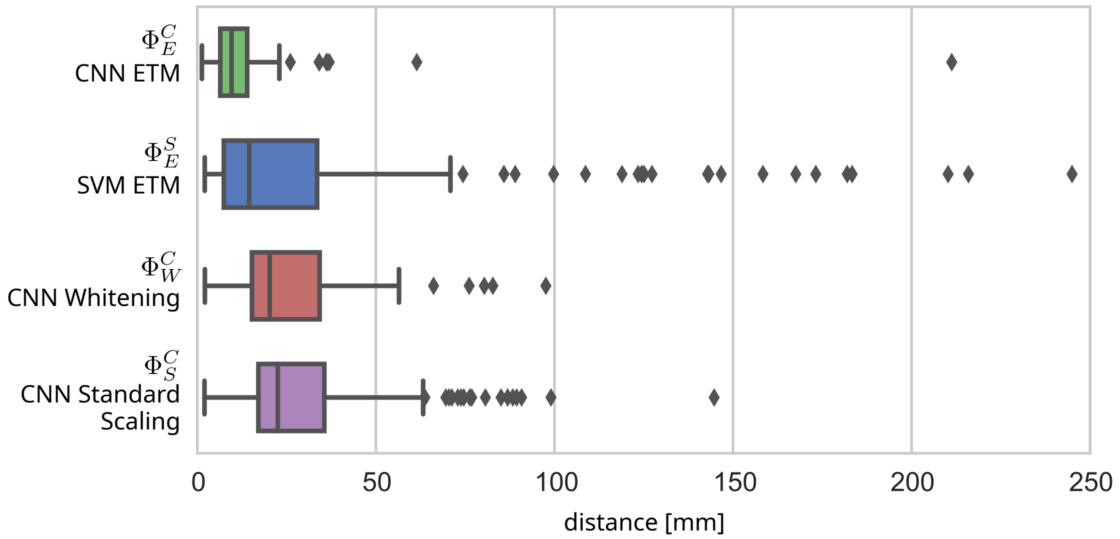


Figure 3.5: Comparison of classifier performances for seed point detection on $\mathcal{D}_{test}$. The distances indicate the distance from the ground truth center to the detected seed.

	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	Total
shift by 1 disc label	1 ↑	1 ↑	-	1 ↑	2 ↓	1 ↑	6
shift by 2 disc labels	-	-	-	1 ↑	-	-	1
shift by > 2 disc labels	-	-	-	-	1 ↑	-	1
missing	-	-	-	-	2	-	2
sacrum seeds in cervical scans	-	-	-	0/1	0/4	-	0/5

Table 3.2: Number of scans with shifts and missing seeds in seed point localization using Φ_E^C. Arrows indicate a shift towards the cervical spine (↑) or towards the sacrum (↓).

and a screw that confuses ETM matching in CT data ($\mathcal{D}_6$). One major shift is observed due to misclassifying parts of the thoracolumbar spine at the scan border as the sacral region in $\mathcal{D}_7$ (see Figure 3.6 (d)).

Our proposed CNN Φ_E^C outperforms the classic SVM and both CNNs with standard preprocessing as they tend to confuse patches with other parts of the spine. From the results in Figure 3.5, we conclude that CNNs in general grasp the basic structure of sacrum patches better than the SVM. Moreover, ETM processing further improves the classification. To further reduce the number of misclassifications, the patch size could be increased to capture a larger part of the spine. Moreover, the classifier could be improved by adding more context, for example, through the use of auxiliar 2D projections as additional input. Training a 3D patch classifier would be another option, yet, computationally more expensive compared to a 2D or 2.5D model.

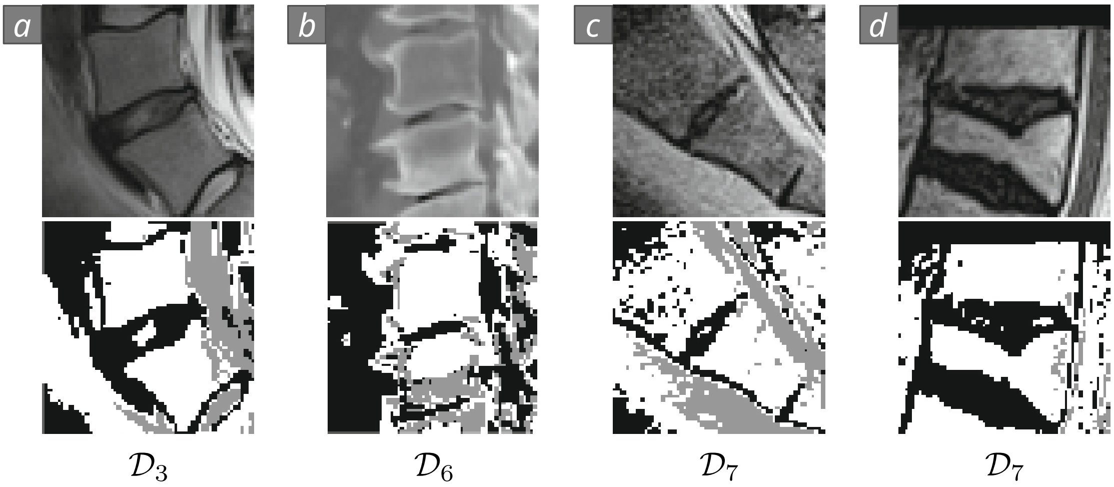


Figure 3.6: Visual examples of misclassified sacrum patches obtained with our model Φ_E^C for different datasets (top row: original patch, bottom row: corresponding normalized patch): (a) shift by one disc label ↑, (b) shift by two disc labels ↑ in pathological CT scan, (c) shift by one disc label ↓, (d) misclassification of thoracolumbar region as sacral region (shift by > 2 labels ↑).

	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$
p_2	20.4	40.9	11.7	14.0	7.4	15.8
p_4	74.5	90.3	48.4	55.8	31.6	34.2
p_6	92.9	97.4	75.4	76.7	57.3	55.0
p_{10}	95.9	100.0	94.9	95.3	78.4	75.0
p_{15}	99.0	100.0	98.9	100.0	94.1	90.8

Table 3.3: p_ϵ measures [%] for every dataset $\mathcal{D}_i \in \mathcal{D}_{test}$.

3.5.2 Labeling and Localization Accuracy

We initialize the labeling of a scan at the corresponding detected seed. From the scans with shifted seed positions, we can label all five cases with one or two shifts towards the cervical spine, in the sense that the center positions are detected correctly, but manual correction of the labels by a radiologist is necessary. The three scans from $\mathcal{D}_7$ cannot be labeled successfully, because model matching diverges. Therefore, we exclude the cervical scans and scans with missing and shifted seeds in the following analysis, resulting in a total of 130 MR and 16 CT scans.

Overall we reach a labeling accuracy of 92.5 % (135/146 scans), and good localization accuracies p_ϵ for all datasets as summarized in Table 3.3. Figure 3.7 shows the effect of position refinement w.r.t. the center accuracies $\bar{e}$. We reduce the localization errors for all datasets except $\mathcal{D}_7$, where a few outliers are introduced with the proposed refinement. In those cases the ETM normalization and consequently the template matching failed.

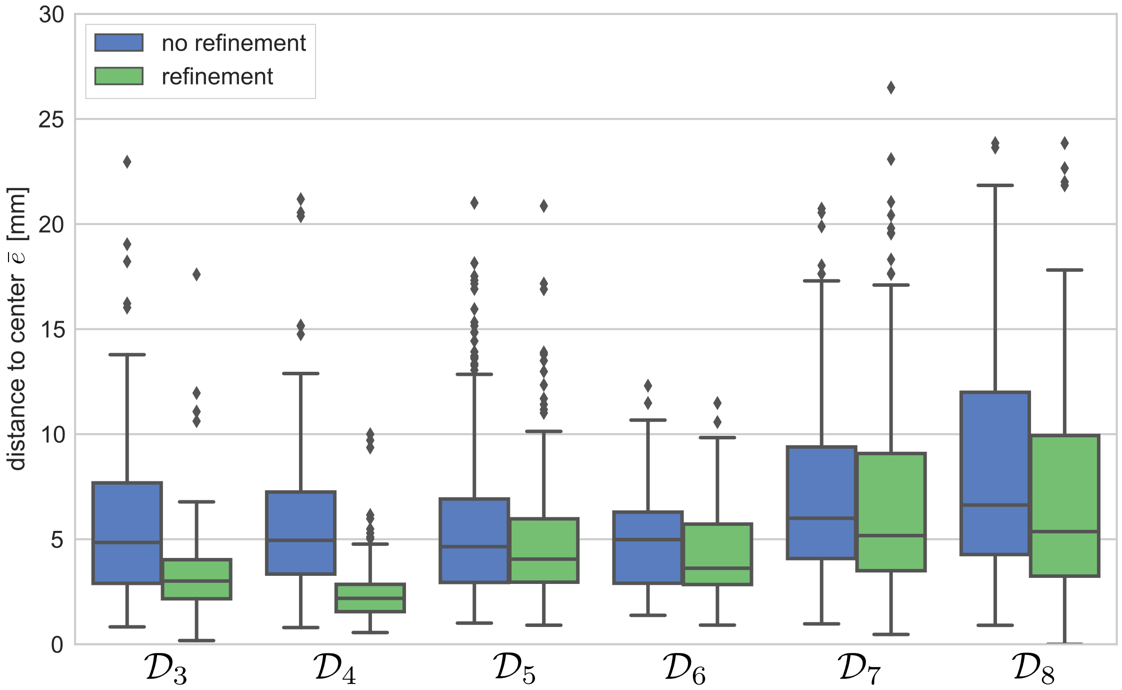


Figure 3.7: Tissue center accuracies $\bar{e}$ per dataset $\mathcal{D}_i \in \mathcal{D}_{test}$ with and without template-based position refinement.

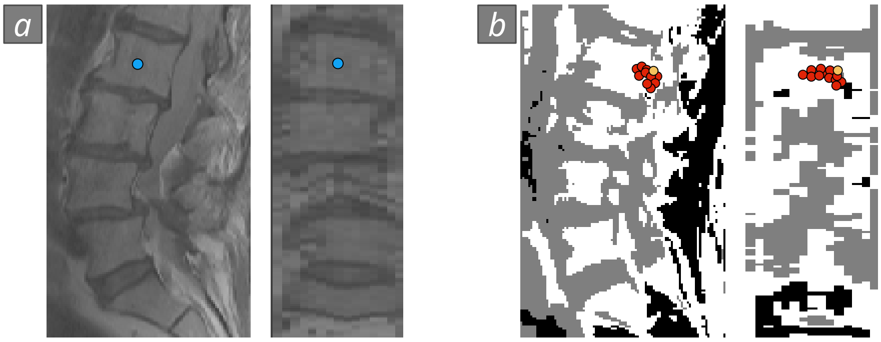


Figure 3.8: Failed refinement of vertebra $L2$ in a scan from $\mathcal{D}_7$: (a) model-matched position for vertebra $L2$ (blue), (b) candidate positions obtained from template matching (red) and the selected, refined $L2$ position (orange). Sagittal (left) and coronal (right) projections are shown for (a) and (b).

Figure 3.8 shows such a case from $\mathcal{D}_7$ where our refinement method results in a worse position for $L2$ compared to the model-matched position. The highest template overlap occurs at a position in the spinal canal due to a non-optimal normalization obtained from model matching.

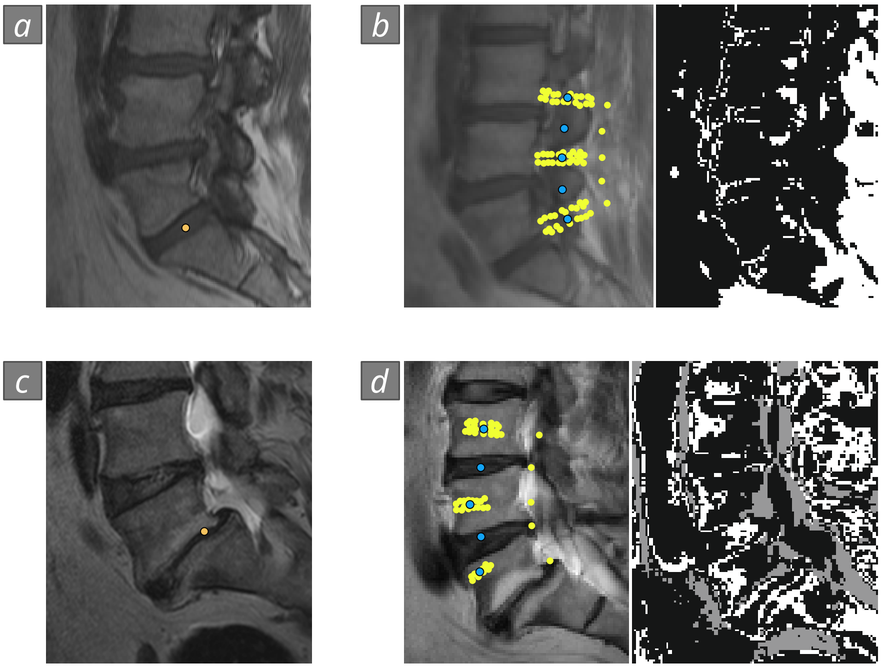


Figure 3.9: Model mismatches in a scan from $\mathcal{D}_7$ (a, b) and a pathological case from $\mathcal{D}_8$ (c, d): (a) and (c) correctly localized $L5/S1$ position. (b) Model-matching diverges into the spinal canal when matching the three-disc model for $L4/L5$. (d) Incorrect model-matching results in a label shift by one label. For (b) and (d), model-matched positions of tissue centers (blue) and remaining model-matched positions (yellow) are shown. Normalized data obtained from model matching is shown to the right.

Visual samples of successful and failed labeling cases from different datasets are provided in Figure 3.10, Figure 3.11, Figure 3.12, Figure 3.13, and Figure 3.14. Labeling fails partially in 11 scans because of model mismatches due to very low image contrast in $\mathcal{D}_5$ and $\mathcal{D}_7$, pathologies ($\mathcal{D}_8$), or initialization at the tissue border ($\mathcal{D}_6$, $\mathcal{D}_7$). Figure 3.9 shows a case from $\mathcal{D}_7$ and a patient with pathologies from $\mathcal{D}_8$ where model matching fails. In both cases, the correctly localized $L5/S1$ position is closer to the border of the spine. This may be a reason why the three-disc model M^{λ_l}, $\lambda_l = L4/L5$ does not match to the correct position. In addition, the scan from $\mathcal{D}_8$ (bottom row) is a spondylolisthesis case, i.e., where one vertebra is displaced compared to another one, as visible between $L5$ and $S1$.

In general, we observe the lowest position errors for $\mathcal{D}_3$ and $\mathcal{D}_4$ and the highest errors for $\mathcal{D}_7$ and $\mathcal{D}_8$ (see Figure 3.7, Table 3.3, as well as the discussion in Section 3.5.3). While for $\mathcal{D}_3$ and $\mathcal{D}_4$ no pathologies were reported and all scans exhibit the same voxel size, $\mathcal{D}_7$ and $\mathcal{D}_8$ are more challenging. Both datasets have a high variation in terms of MR image contrasts, sequences and voxel sizes. Especially $\mathcal{D}_7$ has a very low in-plane pixel size and exhibits the largest variation in slice thickness among all datasets. In addition, $\mathcal{D}_8$ from Zukić et al. [272] comprises mainly pathological cases. Overall, the variation in those

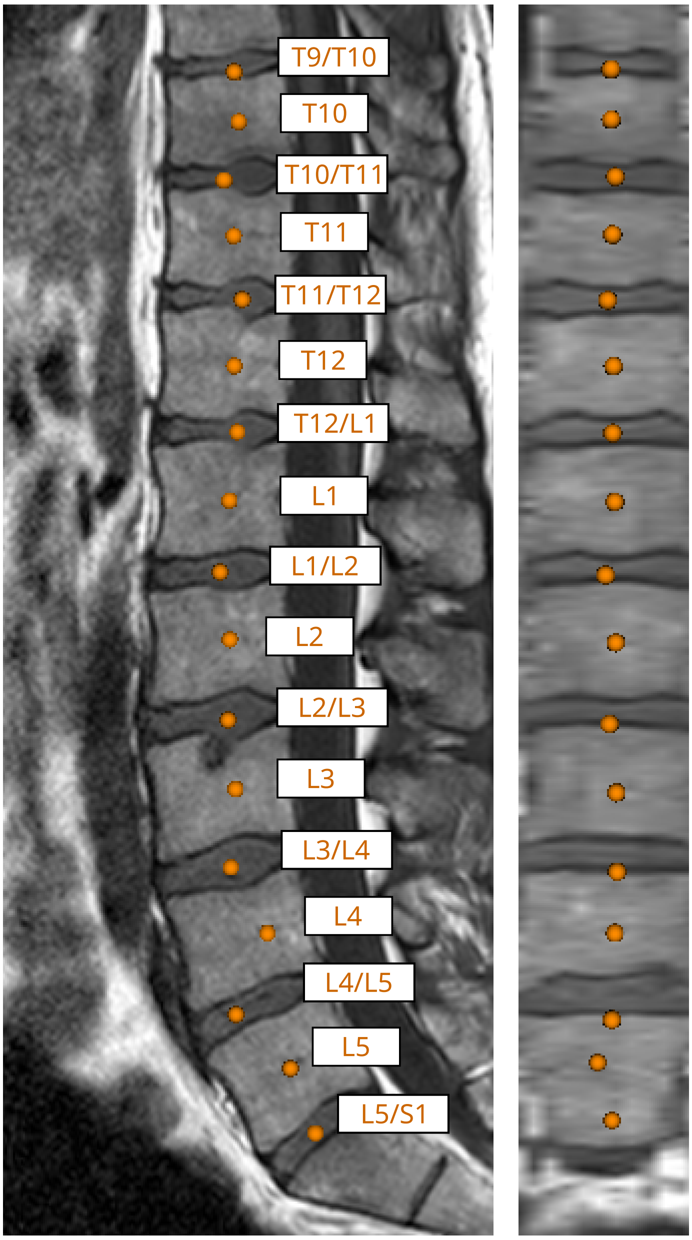


Figure 3.10: Successful labeling of T1w MRI scan from $\mathcal{D}_8$. Sagittal (left) and coronal (right) projections are shown.

two datasets is higher compared to our training sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Increasing the diversity in the training set and adding more (severe) pathological cases could potentially reduce the errors in tissue center positions for more challenging scans.

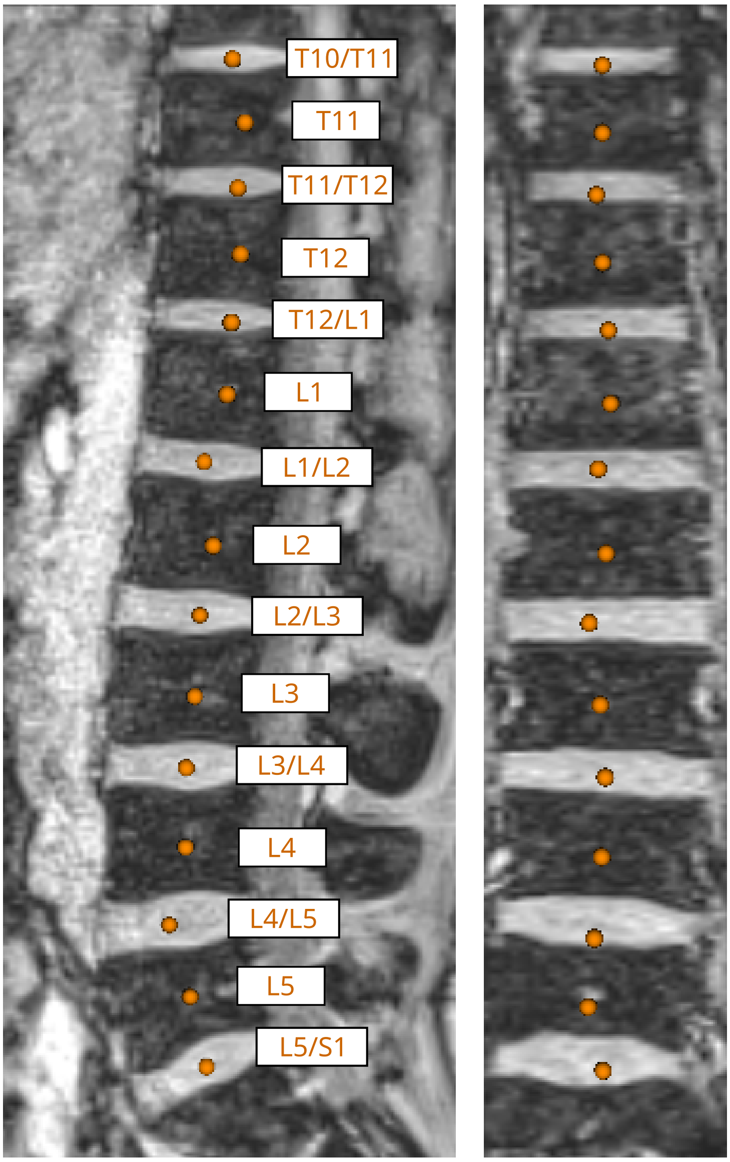


Figure 3.11: Successful labeling of MRI Dixon opposed-phase image channel from a scan from $\mathcal{D}_5$. Sagittal (left) and coronal (right) projections are shown.

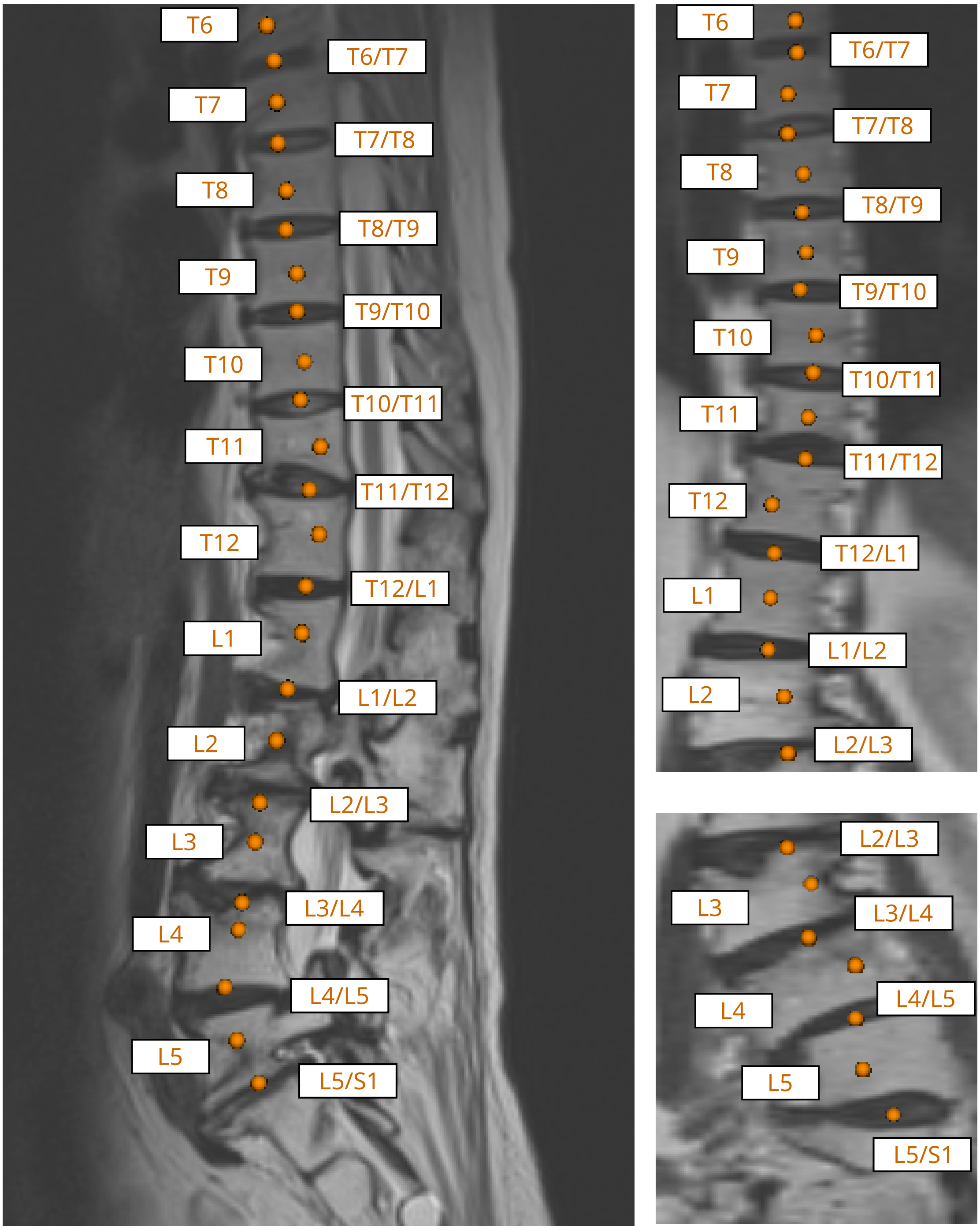


Figure 3.12: Successful labeling of a patient with scoliosis in a T2w MRI scan from $\mathcal{D}_8$. Sagittal (left) and coronal (right) projections are shown.

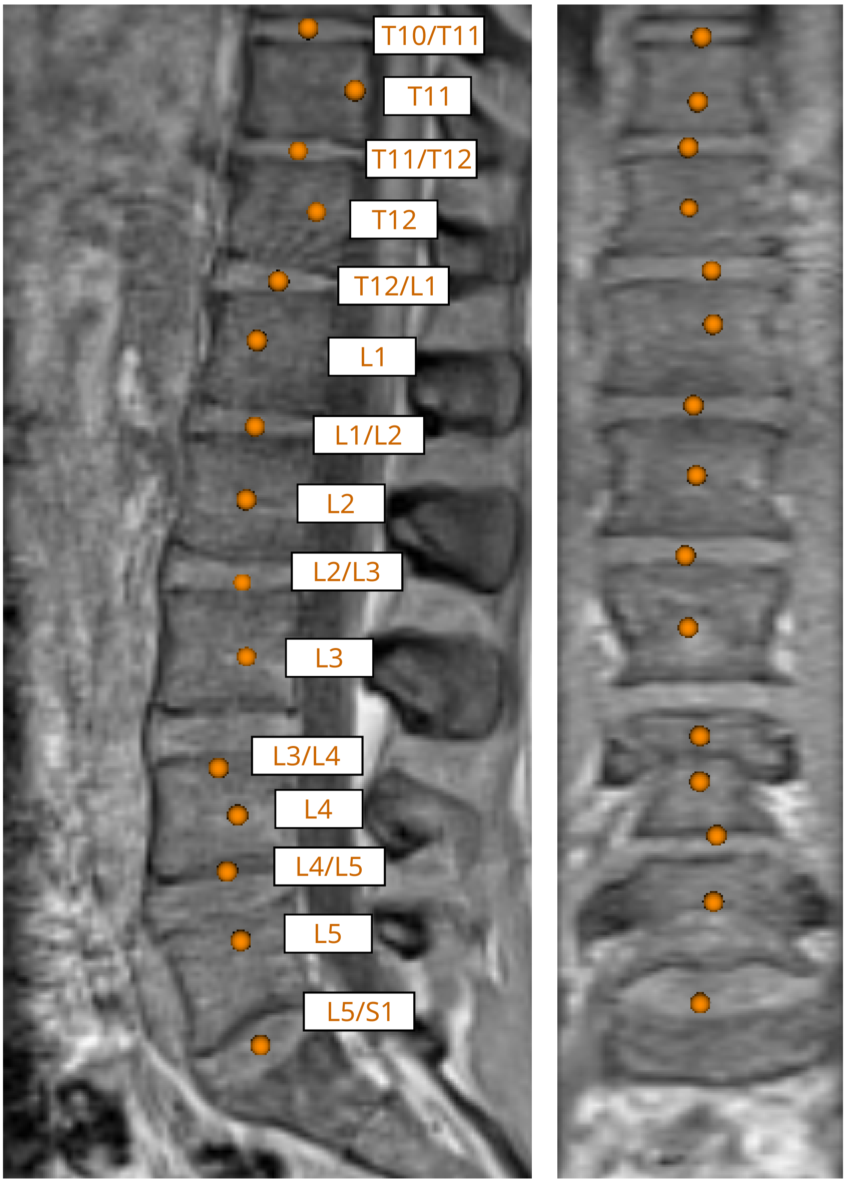


Figure 3.13: Example of a case where labeling failed in a low contrast MRI Dixon inn-phase image channel from a scan from $\mathcal{D}_5$. Sagittal (left) and coronal (right) projections are shown.

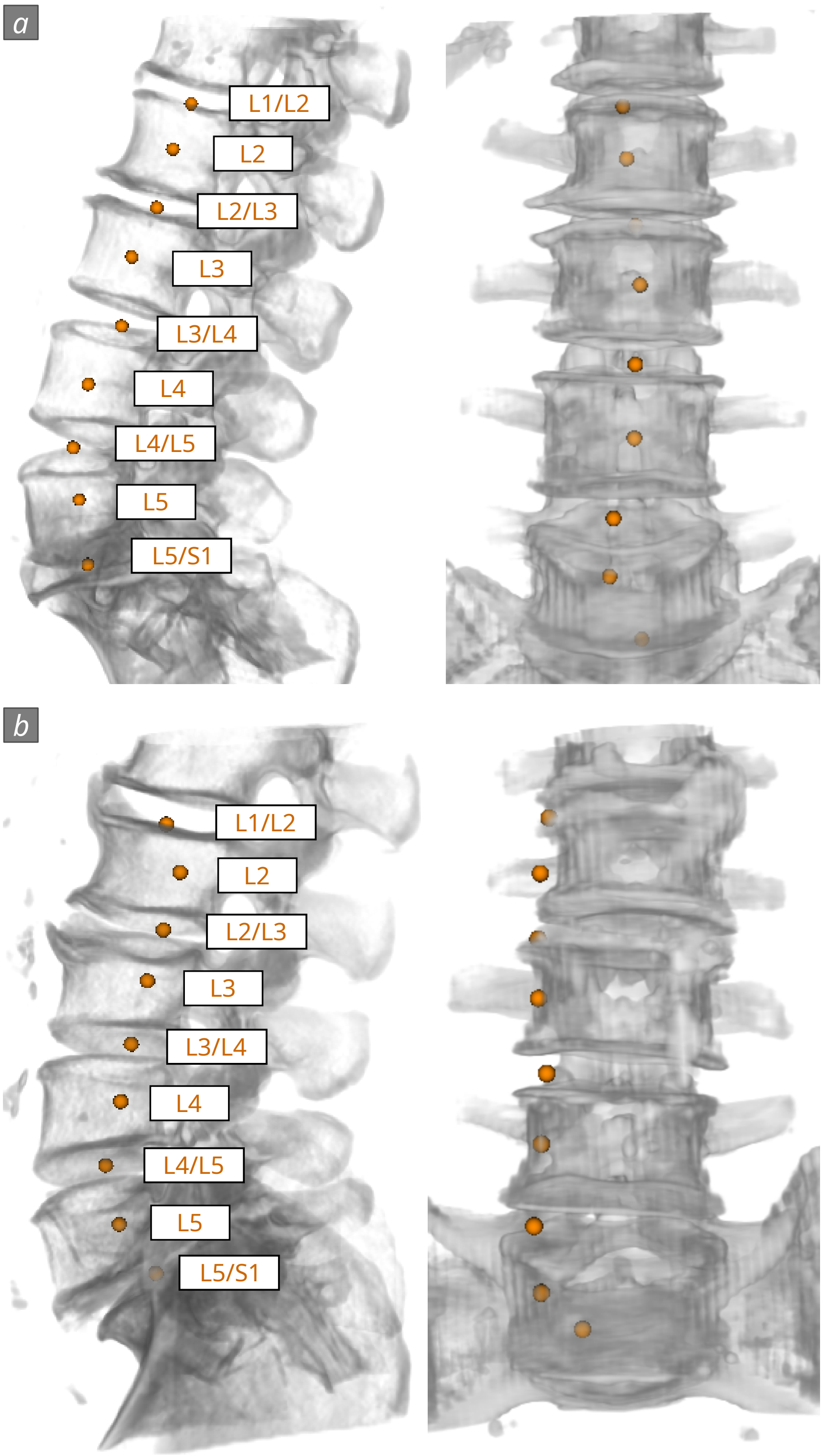


Figure 3.14: Sagittal (left) and coronal (right) projections of spine labeling results on CT scans from $\mathcal{D}_6$: (a) successful, (b) failed labeling.

Method	Distances [mm]	Tissue	Dataset
MICCAI CSI Workshop & Challenge 2015 [1]	[0.9 ± 0.5, 3.9 ± 1.6]	disc	$\mathcal{D}_3$
Jamaludin et al. [96]	1.1 ± 0.6	disc	$\mathcal{D}_3$
Chu et al. [42]	1.6 ± 0.9	vertebra	$\mathcal{D}_4$
Chen et al. [35]	1.3 ± 0.6	disc	$\mathcal{D}_5$
Heinrich and Oktay [79]	3.87 (mean)	disc	$\mathcal{D}_5$
Cai et al. [28]	3.4 ± 2.9 3.1 ± 2.7	vertebra	$\mathcal{D}_6$ 30 scans from $\mathcal{D}_7$
Cai et al. [27]	2.7 to 6.1 2.3 to 5.1	vertebra	$\mathcal{D}_6$ (only 2D image) $\mathcal{D}_7$ (only 2D image)
Štern et al. [220]	2.7 ± 1.9 2.9 ± 1.7	disc and vertebra	CT scans T1w and T2w MR scans
Forsberg et al. [60]	2.4 ± 1.5 2.6 ± 1.6	vertebra	T1w mid-sagittal MR images T2w mid-sagittal MR images
Hojjat et al. [83]	4.54 ± 2.69 4.62 ± 3.19 4.94 ± 2.88 5.12 ± 3.09	disc vertebra disc vertebra	T2w MR T1w MR
Wimmer et al. [247] (semi-automatic)	4.0 ± 3.1 3.7 ± 1.9	disc	$\mathcal{D}_3$ $\mathcal{D}_5$ (without inn-phase images)
Our method	**3.4 ± 2.4** **2.5 ± 1.5** **4.8 ± 2.9** **4.4 ± 2.5** **5.7 ± 3.8** **7.0 ± 5.3**	**disc** **vertebra** **disc** **vertebra** **vertebra** **vertebra**	$\mathcal{D}_3$ $\mathcal{D}_4$ $\mathcal{D}_5$ $\mathcal{D}_6$ $\mathcal{D}_7$ $\mathcal{D}_8$

Table 3.4: Spine labeling results reported in the literature compared to our results.

3.5.3 Comparison to Related Work

Table 3.4 summarizes mean position errors of our proposed approach and related methods on the same or similar datasets. Comparing the performance of our method for all datasets in $\mathcal{D}_{test}$, we report the highest localization accuracies for $\mathcal{D}_3$ and $\mathcal{D}_4$. Both sets include only T2w data, all scans exhibit the same voxel size, and no pathologies are reported [36, 42]. Therefore, Chu et al. [42] and the works presented at the MICCAI CSI Workshop & Challenge 2015 [1] report also very low distances to the annotated centers on $\mathcal{D}_3$ and $\mathcal{D}_4$, often below sub-voxel accuracy. The MICCAI CSI challenge results [1] range from 0.9 ± 0.5 mm to 3.9 ± 1.6 mm on $\mathcal{D}_3$. Jamaludin et al. [96] reach similar results on $\mathcal{D}_3$, but train their method on a different set of T1w and T2w scans. This indicates good robustness of their algorithm. The algorithms evaluated on the Dixon dataset $\mathcal{D}_5$, either fuse features from all image channels [35], i.e., fat saturated, water saturated, inn-phase, and opposed-phase, or use only one image channel [79]. In contrast, our method can be applied independently on each channel. This also applies to our previous work [247] presented in Chapter 2, whereby inn-phase images were not included back then. Li et al. [131] train a multi-scale FCN on $\mathcal{D}_5$ by using a modality dropout strategy, whereby all four image channels are used to label an unseen scan. With this approach, they won the MICCAI CSI Challenge 2016 and report a mean localization error of 0.62 mm on unseen Dixon scans. On the released subset of $\mathcal{D}_8$ by Zukić et al. [272], we obtain correct labeling in 12/14 pathological cases, and in 14/16 cases overall, i.e., in 87.5 % of scans. Figure 3.12 shows a patient with scoliosis, which is labeled successfully. The lower accuracy compared to other datasets is explained by the more challenging nature of the scans. They show high variability in terms of resolution and MR sequences and – except for two scans – comprise of pathological cases only. Zukić et al. [272] perform fully automated vertebrae detection, but the labeling is done in a semi-automatic manner. We note that the authors [272] do not report vertebrae or disc center accuracies, hence, we could not include their results in Table 3.4. Štern et al. [220] also reach high center accuracies similar to other methods in Table 3.4 without re-training of their algorithm, whereby no anatomical labeling is performed. Cai et al. [27, 28] report their results on a subset of $\mathcal{D}_6$ and $\mathcal{D}_7$. Their algorithm requires modality information at test time, which is inferred from DICOM tags. In contrast, our method works completely independent without prior knowledge about the underlying modality. Further, we include the PDw scans in $\mathcal{D}_7$ in our evaluation which are left out by Cai et al. [27, 28]. Excluding those scans, increases our localization accuracy from $p_{10} = 78.4$ % to $p_{10} = 87.4$ % for T1w and T2w scans only. The 2D approach by Forsberg et al. [60] shows good performance on mid-sagittal images but could potentially lead to unsatisfying results, especially in the presence of pathologies like scoliosis. The work by Hojjat et al. [83] also aims at being applicable to various modalities. It does not require previous model training, but user input in a predefined vertebra center.

Comparing again the performance of our spine labeling pipeline to the related works in Table 3.4, we observe lower tissue center accuracies for our method than the state-of-the-art on the same datasets. The higher position errors can be explained by the more general nature of our approach and evaluation setup, as compared to dedicated, data-specific methods.

3.5.4 Generalizability

To further demonstrate the generalization capability of our method, we report additional results by training our models just on CT data $\mathcal{D}_6$ and evaluating them on the remaining MR datasets $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$, $\mathcal{D}_5$, $\mathcal{D}_7$, and $\mathcal{D}_8$.

From the total number of 165 MR scans, 161 show the lumbar spine (see Table 3.1). Seed points are detected correctly in 147/161 scans, i.e., 91.3 % of the cases. This is comparable to the results in Section 3.5.1 where we report a detection rate of 93.6 %. In eleven scans, we observe shifts by one label either towards the sacrum or the cervical spine. In three scans, no seed position could be detected. In general, the mean seed position error is slightly increased by 4 mm on average, compared to our previously reported

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_7$	$\mathcal{D}_8$
p_2	6.5	7.7	30.8	30.5	12.1	11.5	11.0
p_4	30.6	25.3	51.3	59.0	30.1	29.5	31.5
p_6	49.1	48.4	73.1	74.7	50.5	49.6	56.2
p_{10}	71.8	78.0	92.3	85.3	83.9	71.7	83.6
p_{15}	89.5	93.4	98.7	94.7	93.5	86.9	91.8

Table 3.5: p_ϵ measures [%] when training on CT dataset $\mathcal{D}_6$.

seed detection results. Seed points are oftentimes localized closer to the tissue border, which also influences the subsequent labeling. Due to initialization of the three-disc models M^{λ_l} closer to the tissue border, the center accuracies p_ϵ (see Table 3.5) are slightly lower compared to the reported results with MR training data (see Table 3.3). The overall labeling accuracy of 91.8 % (135/147 scans), however, is in line with our findings. Labeling fails for 12 scans where the model matching diverges into the spinal canal or the abdomen.

The lower performance of this evaluation setup is due to the following reasons. First, less training samples are available for the seed localization as well as for ETM model building. Second, the variability within the CT data regarding shape and intensity is significantly lower. While our T1w and T2w MR scans in $\mathcal{D}_1$ and $\mathcal{D}_2$ include several different pathologies (fractures, scoliosis, disc herniation, lordosis, and combinations), the CT data features only cases with minor spondylosis. Thus, we believe increasing the training data and its variability will improve the performance of each building block of the framework.

3.6 Discussion and Conclusion

In this chapter, we presented a novel, fully automatic approach for spine labeling on multiple image modalities using local ETMs. The main advantage and contribution of our method is its generality in terms of imaging variations: First, we can label a wide range of different MR sequences and contrasts, as well as CT scans, without retraining our models. Second, the entropy-objective allows us to apply our pipeline to completely unseen image contrasts and CT scans not covered by training. While many recent works on MR spine labeling do not address the applicability to changing imaging sequences and settings [184], we show that we can apply our method successfully to datasets where just a few scans are available. Dedicated learning methods would overfit on small datasets if retraining is necessary while our cross-modality solution can be applied directly in a fully automatic fashion. This is also an important advantage compared to data-hungry deep learning methods which might not be trainable on datasets comprising only a few scans.

We report promising results on a wide range of publicly available scans for automatic sacral region detection, as well as localization and labeling of intervertebral discs and vertebrae.

Related methods mostly lack proper evaluation on public data. Our results align with recent work. We consider them to compare favorably because we successfully evaluate our algorithm on completely unseen image contrasts not covered by training. Moreover, we report standardized evaluation metrics [267] which we believe is an important step towards comparability of results. We successfully label various pathological scans in $\mathcal{D}_8$ including scoliosis (see Figure 3.12), vertebral fractures, and others. This suggests, that our AAM-based learned entropy models capture shape and intensity variations. They are applicable to the many scans with pathologies resulting in reasonable labeling performance. However, in terms of tissue center accuracies, our proposed general pipeline results in higher position errors compared to the state-of-the-art, as mentioned in Section 3.5.3.

Limitations and Future Work. A limitation of our current method is that the sacrum is required as starting point for the iterative labeling. To the best of our knowledge, the presence of a reliable anchor point, e.g., the sacrum or $C1$ and $C2$ for cervical scans, is a precondition for all labeling algorithms handling MR data, as discussed in Section 3.1.1. We consider the training of additional anchor detectors as future work. In addition, improving the sacrum region classifier to reduce misclassifications is an important topic. This may be achieved, for example, by increasing the patch size or by adding more context from auxiliar projections, as discussed in Section 3.5.1. Moreover, we aim at increasing the center accuracy with, e.g., an improved refinement strategy, or recovery mechanisms in case the labeling fails. Finally, we refer to Section 5.2 for a discussion of the presented method in the context of more recent related work.

CHAPTER 4

Improving Mammography Screening Data Classification

This chapter is based on the following publication:

> Maria Wimmer, Gert Sluiter, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, and Katja Bühler. "Multi-task fusion for improving mammography screening data classification" in *IEEE Transactions on Medical Imaging*, 41(4), pp. 937–950, IEEE, 2022. DOI: 10.1109/TMI.2021.3129068

4.1 Introduction

Breast cancer is the most common cancer type in women and also the leading cause of death by cancer in women worldwide [245]. Fortunately, the mortality rate declined in recent years, one reason being the higher rate of early diagnosis due to the establishment of screening programs. Important cancer risk factors, such as breast density, can be detected and monitored early with such programs [51, 245].

Due to the increasing amount of imaging data, machine learning, especially deep learning algorithms are being developed to automatically process mammography data. Such models perform, for example, localization and classification of lesions [112, 188], breast density classification [102, 123], or cancer risk prediction [154, 255]. These automated methods can be used to accelerate reading workflows [116, 169], or ideally, to support radiologists in image interpretation and diagnosis [21]. Several recent studies report higher accuracies by combining Artificial Intelligence (AI) algorithms with the assessment of a single radiologist [198] or improved performance of radiologists if aided by an AI system [110, 191]. Besides the obtained performance gains, the assistance of radiologists as well as *human-computer collaboration* are becoming increasingly important aspects and challenges for future application in clinical practice [65, 169]. To increase trust in

AI support tools, not only the interpretability of black box models is being intensively studied [22, 187, 208] but also the potential of providing intermediate model results that are linked to radiological features [21, 116]. Recent user studies in cancer screening and diagnosis show that clinicians profit more from models that provide detailed results compared to solutions delivering solely a benign/malignant assessment [25, 228].

4.1.1 Related Work

A standard mammography study is given in Figure 4.1. It comprises four X-ray images that correspond to two different imaging views from each breast: L-CC, R-CC, L-MLO, and R-MLO. Thereby, CC corresponds to the Craniocaudal (CC) view, MLO to the Mediolateral oblique (MLO) view, and L and R indicate the left or right breast, respectively. Radiologists analyze each of the four view images in detail and compare them to obtain a comprehensive view of a patient and render a diagnostic decision [201]. Suspicious lesions, for example, can be visible in one view of a breast but may be obscured in the other view. Therefore, a thorough analysis is necessary. Various deep learning-based methods have been presented in the past years that analyze single- or multiple-view images at a time. However, this is strongly dependent on their task and related clinical question.

The following literature review summarizes related publications published prior to the work [249] on which this chapter is based on. We refer to Section 5.3 for a discussion of more recent works.

Breast Density Scoring

Breast density is an important risk factor as dense breast tissue is related to the development of cancer. Furthermore, Microcalcifications (MCs) and masses are harder to see on the mammograms, causing misdiagnoses [51]. The BI-RADS standard [212] defines density in four categories (a-d) as a measure of the breast tissue composition: “almost entirely fatty” (a), “scattered areas of fibroglandular density” (b), “heterogeneously dense” (c), and “extremely dense” (d). Assessment by radiologists usually has a high inter-observer variability [216] due to the qualitative description of the four categories. Therefore, automated density classification models often focus on the two superclasses “not dense” or “fatty” (a+b) and “dense” (c+d) [102].

Recent works utilize all four mammography views and classify them with multi-view CNNs, following the typical scheme illustrated in Figure 1.8. They are used to classify breast density into the four density categories [251] or in both superclasses [102, 251]. In contrast, Lehman et al. [123] train a ResNet-18 model to classify single mammograms and assign the consensus density across all views for the patient. Another method uses a refined AlexNet to classify the two middle density classes (b+c) [158]. Kallenberg et al. [103] perform unsupervised feature learning from multi-scale patches to segment dense tissue and derive a density scoring per image [103].

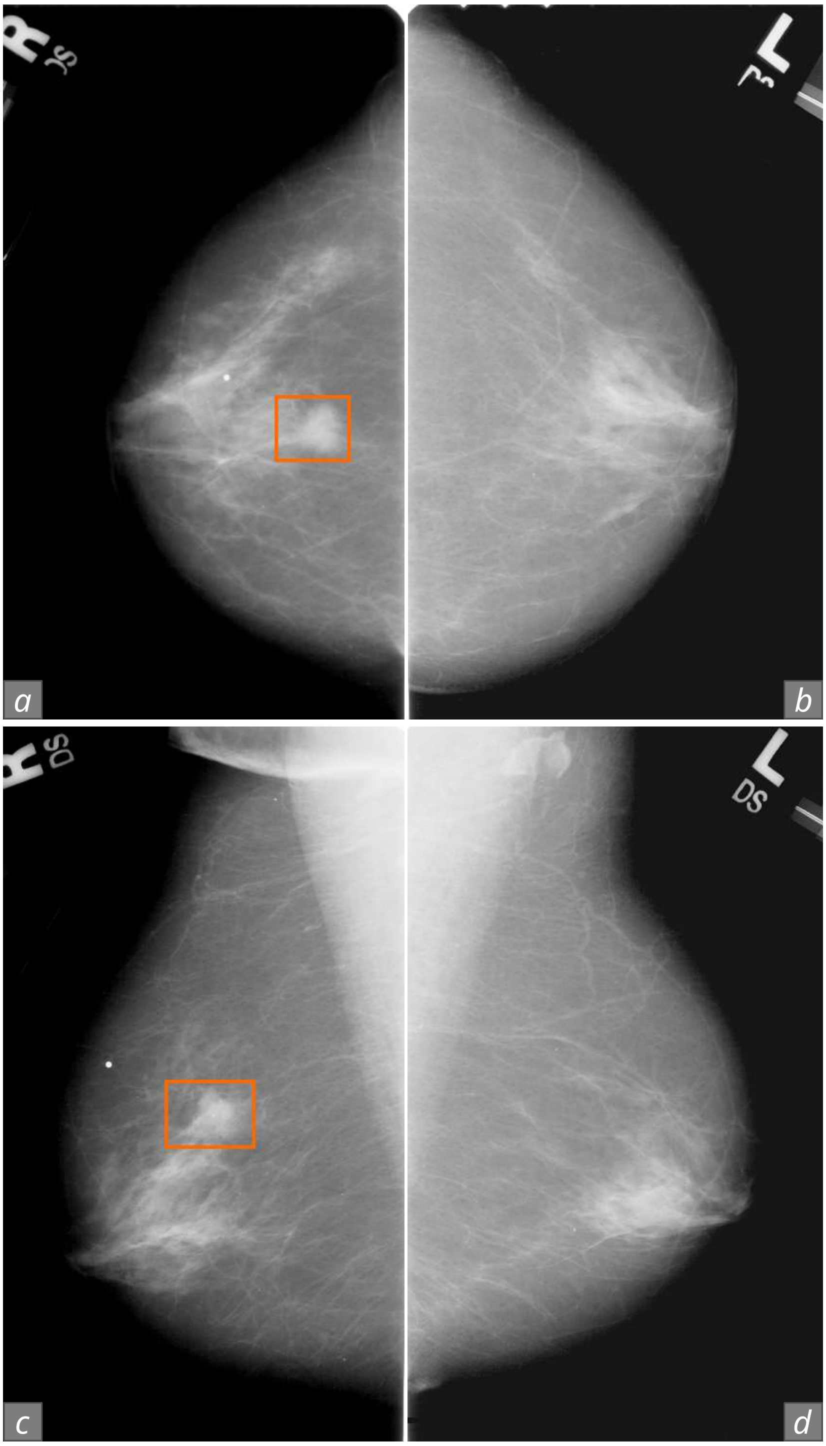


Figure 4.1: Standard mammography study of a patient showing the four standard views: (a) R-CC, (b) L-CC, (c) R-MLO, and (d) L-MLO. The patient has a malignant mass in the right breast, highlighted in orange in (a) and (c).

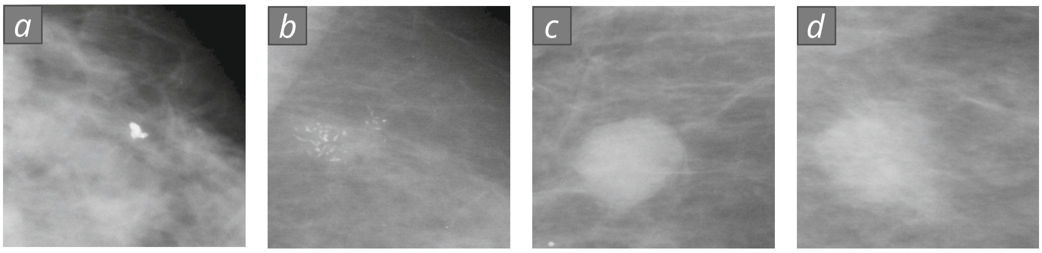


Figure 4.2: Patches showing (a) a benign calcification, (b) malignant MC cluster, (c) benign mass, and (d) a malignant mass.

Lesion Localization and Classification

Exact localization and classification of lesions, i.e., masses, calcifications, and clusters of MCs, in mammograms are crucial as they are important risk factors or already indicators of cancer [201]. Figure 4.2 shows examples for benign and malignant lesions. While many works perform lesion localization, quantification, classification, or all together [9, 30, 111, 112, 188], others solely classify already extracted lesions on patches [15, 21, 52, 160, 194]. The use of classical feature extraction and machine learning methods, or the combination thereof with CNNs, has been intensively investigated in the literature [4, 15, 59, 111, 112]. Mordang et al. [160] were the first to use CNNs for MC localization and utilized a VGG-like architecture for this task. Various studies focus on the classification of MCs and MC clusters [52, 205, 237], e.g., with a combination of a Difference-of-Gaussians detector and two-stream CNNs [237]. Dhungel et al. [53], among many others [4, 11, 15, 21, 194], perform localization and analysis of masses. They combine deep belief networks, Gaussian mixture models, and CNNs for mass detection. A recent work by Barnett et al. [21] proposes an interpretable mass classification framework with the goal of following the reasoning process of radiologists. Finally, state-of-the-art object detection approaches like Faster R-CNN [185] or YOLO [100] have been applied for lesion localization and classification [5, 9, 11, 143, 188]. Ribli et al. [188] utilize a Faster R-CNN with a VGG16 backbone to detect and classify lesions into malignant and benign classes individually. Others extend a Faster R-CNN model by a cascaded classification step to reduce false-positively detected lesions [9].

Malignancy Scoring

Several studies classify single or multiple mammograms directly to obtain a score assessing whether a view image is cancerous [143, 144, 154, 206, 208] or contains a (specific) malignant or benign finding [211, 225, 252, 271]. Recent works utilize, e.g., an all-convolutional design combined with curriculum learning [206], multi-instance learning [143, 207, 208, 271], self-supervised methods [225], or a multi-view-multi-task approach [116]. Wu et al. [252] concatenate heatmaps obtained from sliding window patch classification to classify full images. Other works derive a malignancy score per view image, breast, or patient by averaging or considering the maximum score, e.g., obtained from a Faster R-CNN [188] or a map of pixelwise abnormality scores [110].

Feature or Information Fusion

The fusion of *features* or, more generally, of (extracted) *information* is inspired by how radiologists assess and compare ROIs and mammograms to obtain a comprehensive view of a patient. The term *feature* can refer to "classical, handcrafted" features, e.g., Gabor filters, curvelets, entropy, etc., CNN-features extracted by a CNN, or non-imaging features like patient age. The extraction and fusion is performed at different scales, for example, *locally* from/within a single-view image, patches, or across ROIs [52, 112, 143, 144, 207, 208, 271]. Kooi et al. [112] fuse CNN and classical features extracted from patches within a single mammogram. Lotter et al. [144] and Shen et al. [208] fuse local CNN patch features, whereas the former extract them with a sliding window approach, and the latter extract only CNN features from salient regions obtained with a global image classifier. Another common approach is to utilize *multiple views* for localization and classification of lesions and full images, as summarized by Jouirou et al. [101]. Shachor et al. [205] dynamically combine classical features from local patches from MLO and CC views for calcification classification. Kooi et al. [111] fuse CNN features from ROIs across views for malignant mass detection. The usage of *multi-view CNNs* that follow the basic multi-view architecture, as introduced in Section 1.4.2 (see Figure 1.8), also has been studied [30, 66, 115]. Each view image is processed with a CNN, followed by feature fusion at a given layer. Such models have been trained for different purposes, e.g., BI-RADS scoring [66] or breast density classification [102, 251]. Carneiro et al. [30] use only the CC and MLO views of one breast simultaneously and also include mass and calcification masks as additional input obtained, e.g., via an object detection approach. McKinney et al. [154] propose several models, which use different fusion and combination strategies, e.g., concatenation of CNN patch features across all views, fusion of CNN image-level features per breast and/or patient, or concatenation of non-imaging features like patient age with CNN features.

The last stage is *decision-level fusion*, i.e., fusion of predictions, which has been investigated by Kyono et al. [115], for example. They predict several radiological features, e.g., breast density, diagnosis, age, with a multi-task CNN separately for each view. Next, they fuse the features from all four views, and obtain a benign/malignant classification on a patient level. Finally, the naive ensembling of predictions from different models, e.g., via averaging, can also be interpreted as decision fusion [110, 154, 188].

Summary

While many recent works directly classify ROIs or view images with, e.g., CNNs, a significant part utilizes some form of information fusion to process mammography data. Table 4.1 provides a detailed overview.

Table 4.1: Overview of related works. Type of data (author column): P = Full-Field Digital Mammography (FFDM) processed, R = FFDM raw, U = FFDM unclear, SF = scanned film; Task: C = cancer risk, D = breast density classification, L = lesion localization and/or classification, M = prediction of BI-RADS, benign/malignant, cancer yes/no, etc., on image/patient level; Data: name of image dataset; Fusion: ✓= some form of fusion involved; Intermediate / Sub-results: type of intermediate/additional results provided apart from final scores; Method: brief summary (RF = Random Forests, DBN = deep belief network, GMM = Gaussian mixture model, DoG = Difference of Gaussian).

Author	Task	Data	Fusion	Intermediate / Sub-results	Method
[102] (P), [251] (R)	D	private	✓	no	multi-view CNN
[123] (P), [158] (P)	D	private		no	single-view CNN
[103] (R)	D	private	✓	dense tissue segmentation	multi-scale unsupervised segmentation + texture scoring
[53] (U)	L	INbreast		no	DBN + GMM (localization), CNN + RF (classification)
[15] (SF)	L	BCDR-FM		no	CNN + SVM for classification
[59] (U)	L	private		no	SVM for classification
[4] (SF,U), [194] (SF,R)	L	CBIS-DDSM ([4]), INbreast [4], DDSM [194], private [194]		no	CNN for classification
[160] (R)	L	private		no	CNN for localization + classification
[237] (U)	L	private	✓	no	DoG + multi-scale two-stream CNN
[205] (SF)	L	DDSM	✓	no	multi-view CNN
[52] (SF)	L	DDSM	✓	no	classical features + feed forward network
[112] (R)	L	private	✓	no	candidate localization (RF) + classification (CNN features + classical texture features)
[111] (R)	L	private	✓	no	dual-stream CNN for lesion ROI classification
[21] (U)	L	private	✓	class activation map, mass margin class score	case-based reasoning, compares parts of new images to learned prototypes

Table 4.1 – continued from previous page

[5] (U), [9] (U), [11] (SF), [188] (SF,U)	L	DDSM ([11, 188]), private ([9, 188]), INbreast ([5, 9, 188]), OPTIMAM ([5])		no	Faster R-CNN-/YOLO-based lesion localization + classification
[143] (SF,U)	L, M	DDSM, OPTIMAM, private	✓	benign + malignant lesions (bounding boxes)	RetinaNet-based approach + multi-stage training (fully + weakly supervised, multi-instance learning)
[225] (U)	L, M	INbreast, private	✓	malignancy probability map	self- and weakly supervised reconstruction for lesion localization/segmentation, image-level classification
[30] (SF,U)	M	INbreast, DDSM	✓	no	multi-view CNN
[206] (SF,U)	M	CBIS-DDSM, INbreast		salient regions	all-convolutional CNN (two-stage)
[271] (U)	M	INbreast	✓	no	multi-instance approach
[66] (R), [252] (R)	M	private	✓	heatmaps of malignant / benign + malignant regions	multi-view CNN(s), fusion at different stages
[211] (SF,U)	M	INbreast, CBIS-DDSM		malignant regions	CNN + region-based/global group-max pooling
[110] (U)	M	OPTIMAM, private		pixel-wise abnormality score	semi-supervised CNN (two-stage)
[144] (SF)	M	DDSM	✓	no	multi-scale CNN + curriculum learning
[208] (R)	M	private	✓	saliency maps (malignant findings)	weakly supervised approach, global (weak localization) + local CNN
[115] (P), [116] (P)	M	private	✓	radiological features per view, heatmaps	multi-view, multi-task CNN
[154] (SF,P)	C	OPTIMAM, CBIS-DDSM, private	✓	malignant regions (bounding boxes)	patch-level, image-level and CNN + non-imaging feature fusion (various models)
Our method (SF)	D, L, M	DDSM, CBIS-DDSM	✓	breast density, lesions (bounding box + label), findings classification	task-specific CNNs (multiple scales), feature and prediction fusion with CNNs + MLPs

The reasons for fusion are manifold: it is performed to

(a) incorporate different aspects at different levels (ROI, image, patient),

(b) thus, increase robustness and performance of classification models [52, 111, 112, 143, 144, 205, 207, 208, 271],

(c) and increase explainability and interpretability of model predictions [21, 22, 115, 116, 208].

Methods that fuse predictions across one or more ROIs or mammograms usually build upon models that predict the same scores for the same task or perform standard model ensembling strategies [115, 116, 123, 154]. On the other hand, methods that perform a fusion of features within or across images mostly do not provide intermediate results, e.g., assessment of suspicious regions, but only final classification results. Although recent user studies highlight the potential of providing detailed classification results or pinpointing to suspicious regions [25, 228], only a few proof-of-concept studies explore fusion and the potential of providing intermediate results similar to the assessment of radiologists in the field of mammography [21, 115, 116]. These methods operate only on lesion level [21] or fuse models that predict the same multi-task scores [115, 116]. To the best of our knowledge, the fusion of models trained for *different* tasks is not being studied in the context of mammography.

4.1.2 Contribution

This chapter investigates information fusion for multi-view mammography data from another perspective. We focus on the fusion of features and predictions from *individual, task-specific models* that operate at different *scales* to obtain a comprehensive assessment on a patient level. We address thesis goals **G.4** and **G.5** in this chapter, which relate to the development of methods for *multi-view* and *multi-scale* data and information. We propose a *pipeline approach* comprising

- the development of three *task-specific models*, namely *(i)* a breast density classification model, *(ii)* a lesion localization model, *(iii)* and a findings classifier, as a basis for fusion, and
- the investigation of two fusion strategies: *(i)* the fusion of high-dimensional, task-specific CNN features with a *multi-input embedding CNN* and *(ii)* prediction score fusion of model predictions with MLPs.

By building upon task-specific features and decisions, we obtain *hybrid patient meta-models*, which access the intermediate results in their prediction. Due to the two-stage nature of our method, we report not only a global score on a patient level but make the sub-results that reflect radiological features also accessible to the clinician.

We train both fusion approaches for two different classification targets, which we will refer to as *patient predictions*, i.e., the prediction of the respective model. We predict

- the presence of *any* lesion (*lesion prediction*),
- and whether the patient has any *malignant* lesion (*malignancy prediction*).

At each stage in our pipeline, we aim for resource-efficient models, and therefore, utilize lightweight architectures like MobileNets [86] for image classification-related tasks. The full pipeline is trained and evaluated on the well-known and publicly available Digital Database for Screening Mammography (DDSM) [77, 78] and its curated version, the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [121, 122]. In a comprehensive technical analysis, we show that our task fusion strategy improves patient-level classification over standard model ensembling. A detailed analysis of results and discussion thereof, as well as future clinical perspectives, are provided in Section 4.4.

4.2 Materials and Methods

We start this section by introducing the mammography data we use to train and evaluate our pipeline in Section 4.2.1. Next, Section 4.2.2 explains the three aforementioned task-specific mammography models. They are the basis for our proposed patient meta-models, which we present in Section 4.2.3. We investigate two variants of combining information from task-specific models in a patient meta-model: the fusion of features and the fusion of prediction scores.

Notation: We define a set of four mammography images $\mathcal{I}_i = \{I_i^v\}$ for patient i and mammography image view $v \in \{\text{L-CC, L-MLO, R-CC, R-MLO}\}$. We will refer to this set $\mathcal{I}_i$ as an *exam* or *case* of patient i and to a single mammography image of view v as I_i^v. Further, we define the set of task-specific models $\mathcal{M} = \{B, F, L\}$, where B refers to the breast density model, F is the findings model, and L the localization model. We denote features extracted from an intermediate layer of a task-specific model with $\boldsymbol{h}$ and the final prediction scores of a model with p. Finally, we refer to patient meta-models that fuse information from task-specific models with P. Consequently, a model that performs feature fusion is denoted with P_{feat}, and one that combines prediction scores with P_{score}.

4.2.1 Data

We utilize two publicly available mammography datasets for our experiments: the DDSM dataset [77, 78] and its curated version CBIS-DDSM [121, 122].

DDSM and CBIS-DDSM Dataset

The *original DDSM dataset* [77, 78] comprises 2620 mammography screening exams $\mathcal{I}_i$, collected from four different sites acquired with four different scanners. The data is grouped in four categories:

- "normal" (695 cases): normal exams with no suspicious abnormalities and proven normal exams four years later,
- "benign without callback" (141 cases): cases with benign abnormality but without need for callback,
- "benign" (870 cases): including suspicious findings which were identified as benign findings after callback, and
- "cancer" (914 cases): cancer was proven via histology.

An expert radiologist labeled the breast density per patient and provided pixel-level annotation for abnormalities. Each abnormality is described following the BI-RADS standard [212], including lesion type, i.e., mass or calcification, and further details like shape, lesion margin, and calcification type.

The *CBIS-DDSM dataset* [121, 122] is published at The Cancer Imaging Archive [44] as curated version of the original DDSM set, whereby only images showing one or more lesions have been transferred. Annotated masses were re-checked by a radiologist, and pixel-wise annotations have been refined with an automated segmentation algorithm. However, annotations of calcifications remain unchanged. The authors also provide a predefined split into train and test sets to ensure comparability between methods evaluated on this dataset. Overall, the CBIS-DDSM dataset comprises 3568 annotated lesions (1696 masses, 1872 calcifications) in a total of 3032 mammography view images. For further details on the data, we refer to the original publications [121, 122].

Data Harmonization and Preparation

While providing enhanced annotation quality, the CBIS-DDSM dataset has two shortcomings: first, the absence of normal images without lesions, and second, the lack of full patient mammography exams including all four views. To utilize both resources without losing their individual benefits, we prepare the data as follows:

- First, we preprocess the DDSM set in the same way as it was done for the CBIS-DDSM data, including optical density normalization and remapping the data to the full 16-bit range [2].
- Next, we match, i.e., compare the CBIS-DDSM images to the preprocessed DDSM data to identify corresponding cases and obtain a total of 2590 full mammography exams. We assign the malignancy status of a lesion according to the curated

		Train	Validation	Test	Total
Density	a	207	40	50	297
	b	567	108	176	851
	c	448	86	134	668
	d	289	56	93	438
Pathology	normal	481	107	105	693
	benign	522	98	199	819
	malignant	508	85	149	742
Lesion	normal	481	107	105	693
	mass	545	87	198	830
	calcification	447	90	147	684
	mass & calcification	38	6	3	47

Table 4.2: Number of cases per breast density class, pathology status, and lesion type in train, validation, and test set. 47 scans contain masses and calcifications.

annotation from CBIS-DDSM, whereby “benign without callback” will be treated as a “benign” case.

- Finally, we identify potential ambiguous cases which are originally in the “cancer”, “benign”, or “benign without callback” subset in DDSM but have not been transferred to CBIS-DDSM. Since the status of the lesions for these 329 cases remains unclear, we exclude them. Further, we exclude seven additional exams, which are either incomplete, i.e., not all four views are present, or appear with different imaging data and annotations in different subsets of DDSM and CBIS-DDSM. This leads to our final set comprising *2254 cases*.

Train, Validation, Test Split

We split the dataset into train, validation, and test data on a case level and, thus, ensure that images from one case are not distributed across different sets. We preserve the train/test split of the data provided with the CBIS-DDSM set. The remaining normal cases are randomly distributed in the same ratio (~ 80 % training images) to the train/test set in a way that the distribution of breast density is similar in the three sets. From the obtained train set, we randomly select ~ 12 % of cases for the validation set in a way that the ratio of different breast density classes, lesion types, and pathologies is similar across the three sets (see Table 4.2). Overall, the train, validation, and test set comprise 1511, 290, and 453 cases, respectively. Out of the 2254 cases, 47 cases contain masses and calcifications. In total, 174 cases have more than one lesion, with the maximum number of lesions per case being 24.

4.2.2 Task-Specific Mammography Models

The first stage in our pipeline is the development of a set $\mathcal{M}$ of three resource-efficient, task-specific models $\mathcal{M} = \{B, F, L\}$, which are the base for our patient meta-models P:

- B performs breast density classification on a patient level,
- F predicts the presence/absence of lesions in an image, and
- L delivers bounding boxes around localized lesions and their respective class label in an image.

From each task-specific model, we obtain one or more prediction scores p as output, if applied to an input image or exam. Further, we can extract features, also referred to as feature representations, $\boldsymbol{h}$ at intermediate layers from a model. Both the scores p and features $\boldsymbol{h}$ are used as input to our patient meta-models, which we introduce in Section 4.2.3.

Breast Density Model B

Radiologists include all four view images, i.e., the exam $\mathcal{I}_i$, in the assessment of a patient's breast density. Recent deep learning based density classification models follow this standard and utilize all views as input [102, 251], whereas the usage of only one view has also been studied [123]. We propose a two-stage approach where we employ both ideas in the design of density model B to increase robustness and classification performance.

We build a view model D first that uses a mammography image I_i^v as input to predict one of the two density superclasses, i.e., "fatty" or "dense". The model is built upon a MobileNet classifier [86] with global average pooling, followed by a 1×1 convolutional layer. The architecture is illustrated in Figure 4.3. Our final density model B takes the exam $\mathcal{I}_i$ comprising all four mammography view images I_i^v as input. Each I_i^v is passed to a separate branch, as visualized in Figure 4.4. Each branch consists of the view model D without the last layer, whereby the dropout rate is increased from 0.001 in model D to 0.5 in B. After the following flattening operation, the four 1D feature vectors are concatenated. As we utilize the original MobileNet [86] architecture as feature extractor, each 1D vector has a length of 1024, i.e., the concatenated representation is 4096-dimensional and is denoted as $\boldsymbol{h}_B$ in the following. The final dense layer predicts the density superclass. Applying density model B to an exam $\mathcal{I}_i$ results in the predicted scores for the classes "fatty" and "dense" at the patient level. We refer to the prediction corresponding to the class "dense" as density score, which we denote with p_B.

Findings Model F

The objective of the findings model is to classify a single view images I_i^v into "normal" or "image containing *any* findings", i.e., lesions. Such a model could, for example, be

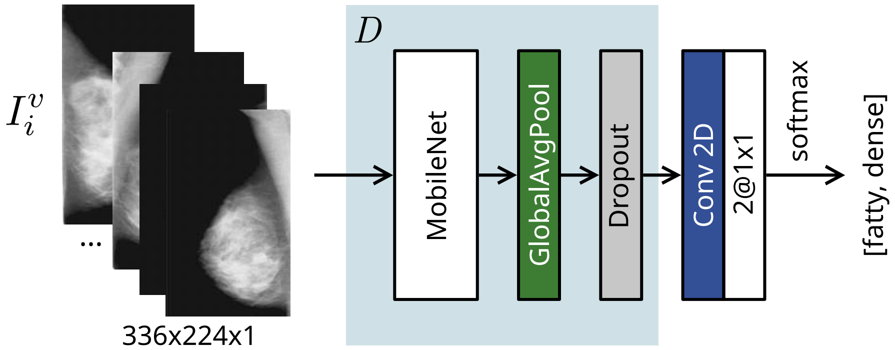


Figure 4.3: The architecture of the density view model D is based on a MobileNet [86] classifier. D takes a view image I_i^v as input and predicts whether the breast tissue is "fatty" or "dense".

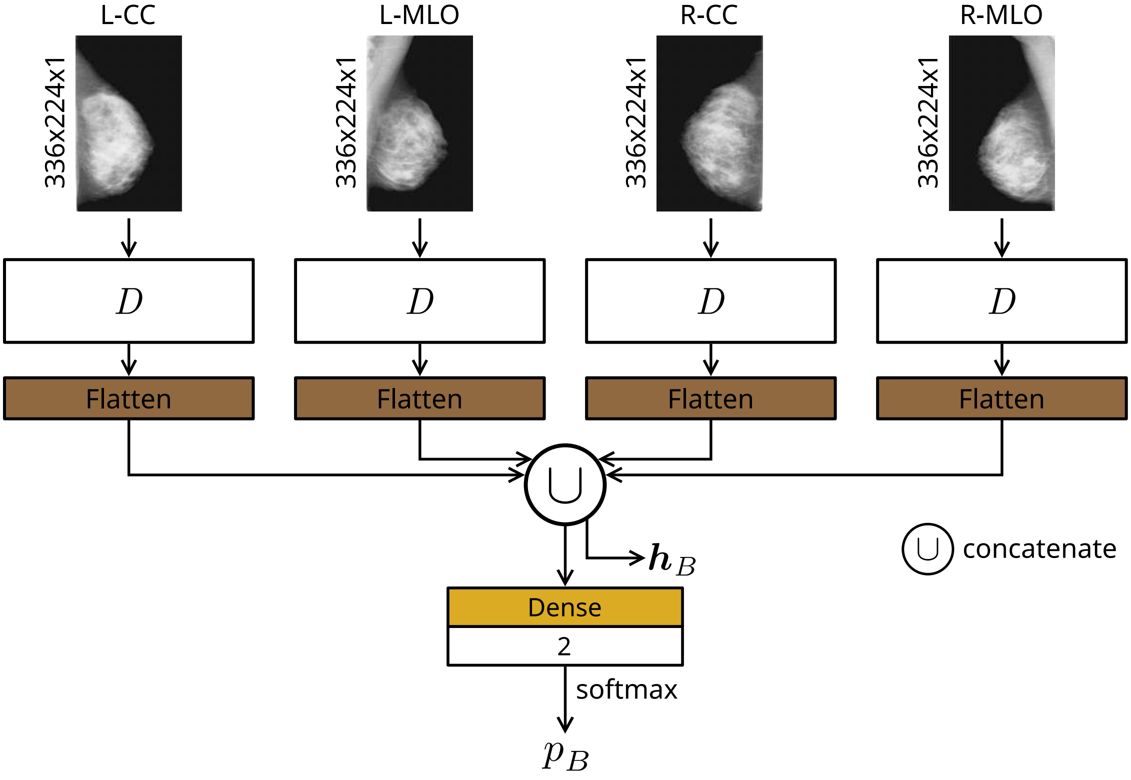


Figure 4.4: Breast density model B takes the four mammography view images I_i^v as input in separate branches, each consisting of a view model D without the last layer. The model's output is the density score p_B, i.e., the prediction score corresponding to the "dense" class. $\boldsymbol{h}_B$ is the concatenated 4096-dimensional feature vector.

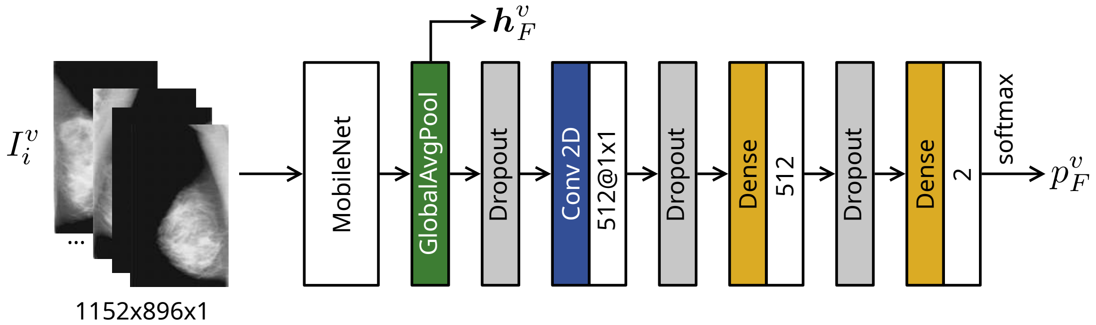


Figure 4.5: The findings model F utilizes a MobileNet[86] feature extractor, followed by alternating dropout and dense layers. F uses a view image I_i^v as input and predicts the score p_F^v which denotes if there is a lesion in I_i^v. $\boldsymbol{h}_F^v$ is the 1024-dimensional feature vector extracted after global average pooling.

integrated into a reporting system, where images with lesions are examined first by a medical expert. Again, we aim for a resource-efficient model to solve this task. We extend on our previous work [125, 151], where we already successfully apply MobileNet [86] in this context. Figure 4.5 illustrates our findings model F with a MobileNet feature extractor and a modified classifier on top as compared to the original MobileNet architecture. Adding an additional dense and dropout layer increases the classification accuracy and the generalization capability of the model. Additionally, we use an increased dropout rate of 0.5 to stronger regularize the network. The output of F for a view image I_i^v is the score p_F^v, which determines whether there is a lesion in I_i^v. Further, we extract the features after the global average pooling layer and flatten them. Again, this results in the 1024-dimensional feature vector denoted as $\boldsymbol{h}_F^v$.

Localization Model L

Similar to radiologists, we aim to detect the exact location of lesions within an image I_i^v and classify them into their correct type and malignancy status. The localization and characterization of lesions are important tasks, as they can be risk factors or already indicators of cancer [201]. Therefore, we develop model L to localize lesions and classify them in either “benign calcification”, “malignant calcification”, “benign mass”, or “malignant mass”. Inspired by recent works on lesion localization [5, 9, 188], we utilize the well-known Faster R-CNN [185] architecture. InceptionV2 [222] serves as feature extractor, which is already successfully applied in the context of mammography lesion localization [5]. Figure 4.6 illustrates the architecture. We propose to use a larger input size for the view image I_i^v as compared to models B and F to not miss tiny structures such as calcifications. Our localization model L classifies localized lesions into the four mentioned types and assigns a score $p_L^{v,k}, k \in [1, n]$ to each of the n detected lesions in I_i^v. Further, we obtain a corresponding 1024-dimensional representation $\boldsymbol{h}_L^{v,k}$ after ROI pooling in the network.

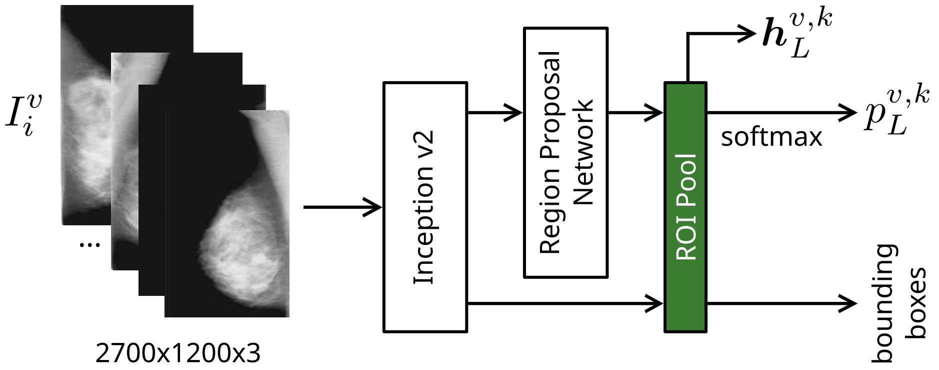


Figure 4.6: The localization model L has a Faster R-CNN [185] architecture with an InceptionV2 [222] feature extractor. The model uses a view image I_i^v as input and predicts a bounding box and the corresponding class label and prediction score $p_L^{v,k}$ for every localized region. Further, we extract corresponding features $\boldsymbol{h}_L^{v,k}$ after ROI pooling.

4.2.3 Patient Meta-Models P

We develop hybrid patient meta-models P as a second step in our pipeline. The meta-models aim to efficiently combine, i.e., fuse, the set of task-specific models $\mathcal{M}$ to obtain a comprehensive patient-level assessment while preserving the individual model predictions related to radiological features and risk factors.

The fusion of models can be performed at various stages, as already summarized in Section 1.4.2. Again, our goal is to develop *resource-efficient* variants. For this, we compare two different fusion strategies:

- late fusion, i.e., the fusion of *prediction scores* p from task-specific models $\mathcal{M}$, and
- intermediate fusion, i.e., the fusion of *features* $\boldsymbol{h}$ extracted from intermediate layers of the individual models.

We denote a model that performs prediction score fusion with P_{score} and one that fuses features with P_{feat}. The models P_{score} and P_{feat} are trained for two different classification targets, which we refer to as *patient prediction*. The obtained patient-level prediction score is indicated with p_P. We consider the following two patient predictions:

- *lesion prediction:* whether the patient has any lesion, regardless of pathology, and
- *malignancy prediction:* whether the patient has any malignant lesion.

Fusion of Predictions with P_{score}

The three task-specific models in $\mathcal{M}$ deliver different prediction scores p at various levels, i.e., at the patient, image, or ROI level. In P_{score}, we concatenate these predictions of the models introduced in Section 4.2.2 to form the vector $\boldsymbol{w}$, formally:

$$\boldsymbol{w} = p_B \cup p_F^v \cup p_L^{v,k} \tag{4.1}$$

with $k \in [1, n]$ and n being the number of considered detected lesions per view. In case of no detected lesions by model L or less lesions than specified by n are found, a probability of 0 is assigned, i.e., $p_L^{v,k} = 0$. For the malignancy prediction, only scores $p_L^{v,k}$ corresponding to malignant masses and calcifications are considered in the combined scores vector $\boldsymbol{w}$. In case no malignant lesions or less malignant lesions than specified by n are found, we set $p_L^{v,k} = 0$.

Fusion of Features with P_{feat}

Apart from the fusion of prediction scores p, we also propose the fusion of feature vectors $\boldsymbol{h}$ extracted from the three different task-specific models in $\mathcal{M}$ with patient meta-model P_{feat}. Therefore, we extract features at the following stages in the networks:

- $\boldsymbol{h}_B$ is the 4096-dimensional, flattened, concatenated representation of view features from B, i.e., the concatenation of features extracted from each view branch after the respective global average pooling layers,
- $\boldsymbol{h}_F^v$ is the 1024-dimensional representation for view image I_i^v, obtained after global average pooling in F,
- $\boldsymbol{h}_L^{v,k}$ is the 1024-dimensional representation for a detected lesion k in I_i^v after ROI pooling in L.

We propose an embedding network that takes the extracted, high-dimensional feature representations $\boldsymbol{h}$ as input in separate branches. Each branch corresponds to one task-specific model. The architecture is illustrated in Figure 4.7. Each channel, i.e., the last dimension in each input branch, corresponds to the respective features of a view image I_i^v. The density and findings branches consist of two convolution blocks, followed by pooling operations. The localization branch utilizes an additional convolution and pooling block for better feature learning. Before and after concatenation of all feature representations, we perform ReLU activations. The final classification part of the network consists of two dense layers with an intermediate dropout layer with a dropout rate of 0.1, followed by a final softmax activation.

Again, we vary the number of lesions considered per view $n \in \{1, 2, 3, 4, 5\}$. In case no lesions are detected with model L, or less lesions than specified by n, background features are pooled from the feature map and used as input. For the malignancy prediction, only features $\boldsymbol{h}_L^{v,k}$ corresponding to malignant masses and calcifications according to the localization model L are considered for the feature fusion. In case of no malignant lesions or less than specified by n, again background features are considered as model input, as we require n lesions per view.

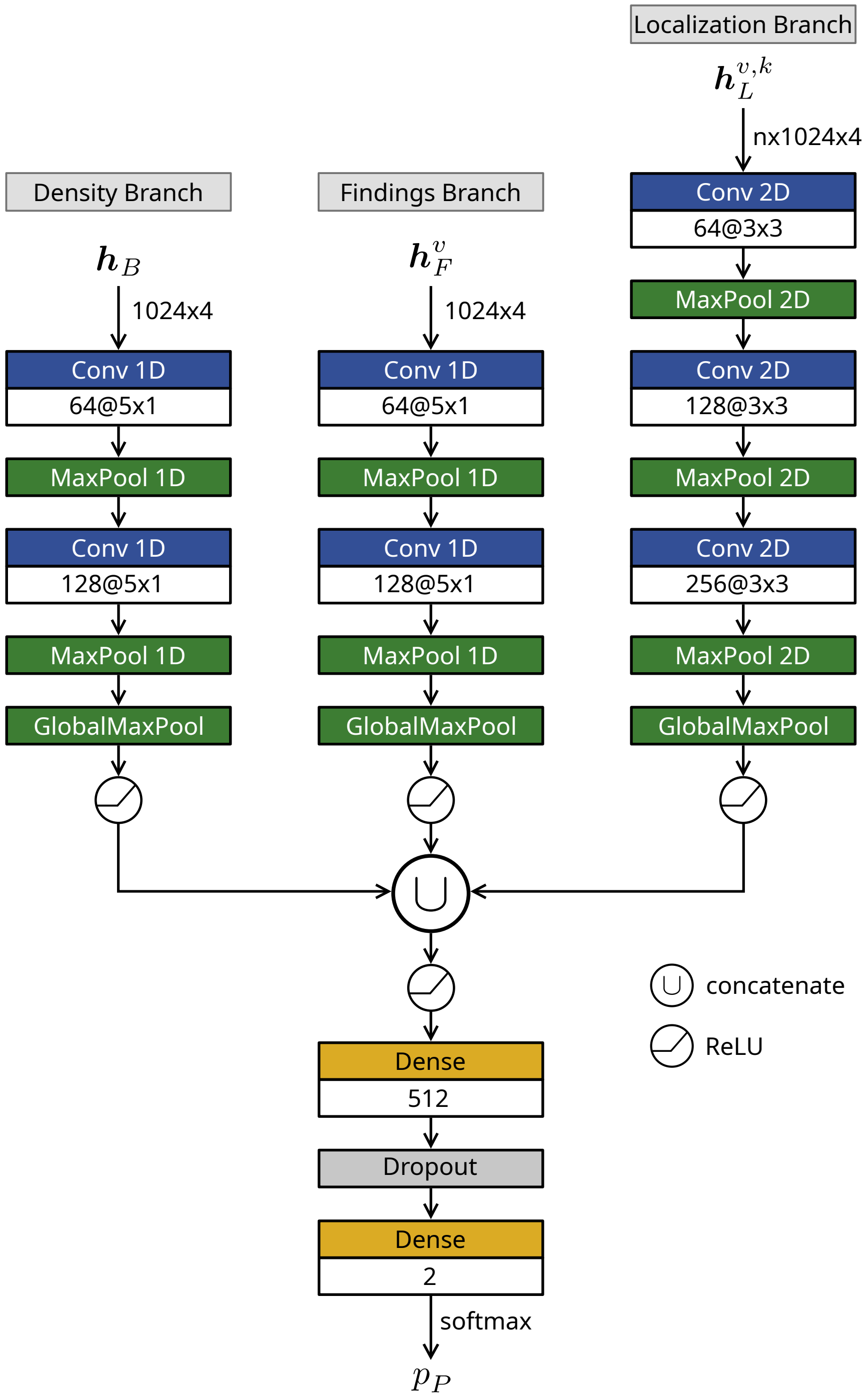


Figure 4.7: Architecture of patient meta-model P_{feat}: It takes the extracted features from the set of task-specific models in $\mathcal{M}$ as input in three separate branches. Each channel in the respective inputs (last dimension) corresponds to the features of a view image. The output p_P of P_{feat} denotes the patient-level prediction score and represents either the lesion or malignancy prediction.

4.3 Experimental Setup

Our framework is implemented in Python, utilizing Keras [41] with the Tensorflow backend [3] for training the task-specific models B, D, F, L, and patient meta-model P_{feat}. Additionally, we use the Tensorflow Object Detection API [88] to train localization model L and scikit-learn [175] for training patient meta-model P_{score}. Model training and experiments are conducted on an NVIDIA Titan X GPU with 12 GB RAM. The storage requirements of trained models range from less than 1 MB for P_{score} to 7 MB for P_{feat}, 47 MB for findings model F, 50 MB for localization model L, up to 75 MB for the breast density model B. Inference for a patient, i.e., for exam $\mathcal{I}_i$, is done within seconds for the complete pipeline, as the inference time for each individual model is less than 1 second.

4.3.1 Training Details

For the training of every task-specific model, we first segment the breast with a basic, non-learning-based segmentation approach according to Shen et al. [206]. Segmentation of the breast has been frequently used by related works as first preprocessing step, e.g., to clean/remove the image background, i.e., the non-breast area in a view image, or for subsequent cropping to the breast area [206, 211, 225, 271]. Similarly, we use the obtained binary mask indicating the breast area to clean the image background and for the sampling of patches inside the breast for pre-training the findings model F (see below). The following set of random data augmentations is executed in each model training:

- horizontal flips,
- rotations in the range $[-15, +15]$ degrees,
- and random sized crops in the range $[85\%, 100\%]$ of the image size.

All image resizing operations are performed using bicubic resampling.

Breast Density Model B

All images are resized to $336 \times 224 \times 1$ with rescaled intensities to the range $[0, 255]$ in floating-point precision to preserve the bit depth. Since the complete breast tissue is of interest for breast density classification, we use a smaller input image size for B as compared to F and L, as the two superclasses are separable also in the smaller resolution. Model training is performed in a two-stage approach with the Adam optimizer and cross-entropy loss: First, imagewise pre-training of the view model D (see Figure 4.3) is performed for 25 epochs and an initial learning rate lr of 1e-3. Further, we employ Stochastic Weight Averaging (SWA) [94] with an initial epoch of 10 to increase the generalization capability of the model. In addition to the standard set of augmentations, random shears are applied. Second, we train the patient-wise model as shown in Figure 4.4. Each view branch is initialized with the SWA-weights from Stage 1, and the complete

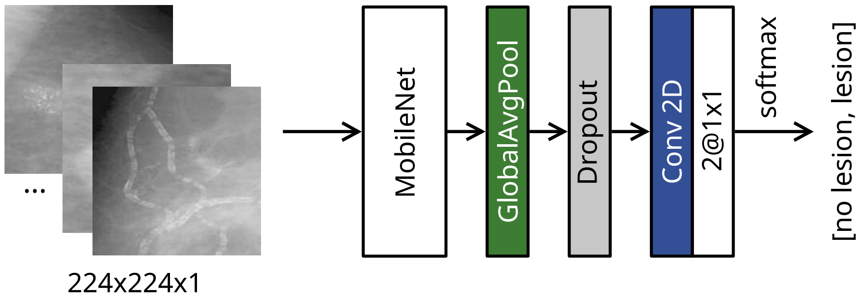


Figure 4.8: The architecture of the patch model used for pre-training the findings model F follows a standard MobileNet [86].

model is trained for 25 epochs (lr = 1e-4). SWA is used with an initial epoch of 5. Horizontal flipping is not performed to preserve the original position of the breast in each view, but instead blurring and grid distortion are additionally carried out to further regularize the model training. We reduce the learning rate by a factor of 0.2 with a patience of 5 epochs on the validation loss in both training stages.

Findings Model F

We perform two-stage training of the findings model F, a strategy already successfully applied by recent works [110, 143, 206]. In both stages, the models are optimized using the Adam optimizer with cross-entropy loss. First, we train a MobileNet-based patch classifier (see Figure 4.8) from scratch with patches of size $224 \times 224 \times 1$, inspired by Shen et al. [206]. We extract an initial set by sampling 5 patches per lesion with an overlap $> 90\ \%$ with the lesion, and 5 patches from normal images with an overlap $> 90\ \%$ with the breast. The patch model is trained with a batch size of 64, lr = 1e-4, and early stopping on the validation loss (patience = 10 epochs, tolerance = 0.001). We perform the following additional augmentations to further increase the diversity of patches: vertical flips, transposes, and shifts/scales/rotations. The model is fine-tuned in a second training iteration with a reduced learning rate of 1e-5.

In the second stage, we initialize the feature extractor of the findings model F illustrated in Figure 4.5 with the obtained patch weights. The full images are resized to $1152 \times 896 \times 1$, following Shen et al. [206], rescaled to $[0, 1]$ and z-score normalized. F is trained using a batch size of 6 with lr = 1e-4. As opposed to the patch model, the validation AUC score is monitored as the criterion for early stopping (patience = 10 epochs, tolerance = 0.001). To further improve the generalization capability, we use SWA with an initial epoch of 5. The model is fine-tuned in a second training round with lr = 1e-5. In both training iterations of model F, we additionally augment the data by vertical flips. Further, we perform stratified sampling utilizing the imbalanced-learn library [124] to balance batches between images showing lesions and normal images.

Localization Model L

The InceptionV2 [222] backend is initialized with COCO-weights and then fine-tuned for the mammography lesion localization task for the four classes. The ground truth bounding boxes required to train the Faster R-CNN model are derived from the pixelwise annotated lesions. We consider the axis-aligned minimum bounding box which encloses the lesion. Only images with at least one lesion are included in the training. We resize the view images to 2700 × 1200 to preserve the small structures of interest and train L with stochastic gradient descent (momentum $= 0.9$, lr = 1e-4) for 100k iterations and a batch size of 2. In addition to the default data augmentation strategies, bounding boxes are randomly jittered with a ratio of 0.005.

Patient Meta-Models P

We perform a parameter search over the number of considered lesions $n \in \{1, 2, 3, 4, 5\}$ for P_{score} and P_{feat} and train all models according to the predefined data split for the lesion and malignancy prediction. Best models are selected based on validation AUC score and recall.

Fusion of Predictions. Prediction scores are concatenated according to Equation 4.1 to obtain one feature vector $\boldsymbol{w}$ per patient. We vary the number of detected lesions $n \in \{1, 2, 3, 4, 5\}$ considered per view and include only their scores. For comparison, we train a classic SVM with RBF kernel, an MLP, and a Random Forest classifier. We perform a parameter search over the following parameters of the individual models and select those with highest validation AUC:

- SVM RBF: $C = \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 500, 1000\}$,
- Random Forest: number of trees $= \{3, 5, 7, 10, 15, 20\}$,
- MLP: layer configuration $= \{[|\boldsymbol{w}|, 2], [|\boldsymbol{w}|, |\boldsymbol{w}|, 2], [|\boldsymbol{w}|, |\boldsymbol{w}|/2, 2]\}$.

Fusion of Features. Before feeding the feature representations to P_{feat}, they are normalized with ψ, where $\psi : \mathbb{R}^n \mapsto [-1, 1]$, resulting in normalized representations $\psi(\boldsymbol{h}_B)$, $\psi(\boldsymbol{h}_F^v)$, and $\psi(\boldsymbol{h}_L^{v,k})$. We optimize P_{feat} with the Adam optimizer using cross-entropy loss and a batch size of 8 and $lr = 5\mathbf{e}\text{-}4$. Early stopping is used with a patience of 10 epochs on the validation loss (tolerance $= 0.001$). Again, we balance batches to ensure equal distribution of classes.

4.3.2 Evaluation Metrics

We compare the performance of classification-related tasks B and F by calculating widely used metrics in the field: the True Positive Rate (TPR), also referred to as *sensitivity* or *recall*, the True Negative Rate (TNR), also referred to as *specificity*, accuracy, and F1-score (F1), i.e., the harmonic mean of precision and recall. Further, we calculate the

Area Under Curve (AUC), i.e., the area under the Receiver Operating Characteristic (ROC) curve, which shows the TPR against the false-positive rate (1 - TNR). Additionally, we provide the Area Under the Precision-Recall Curve (AUPRC) for comparisons with recent studies [115, 205, 208]. For the localization model L, we show FROC curves to measure its detection performance and calculate the number of False Positives per Image (FPI). We report the TPRs for a given number of FPI and refer to this as TPR @ FPI.

4.4 Results and Discussion

This section summarizes intermediate results obtained with task-specific models in Section 4.4.1 and final predictions from score and feature fusion in Section 4.4.2. Section 4.4.3 summarizes ablation study results, and finally, Section 4.4.4 provides an in-depth discussion and analysis of the presented results.

We performed Wilcoxon signed-rank tests on the predictions for task-specific models, fusion models, as well as for ablation studies. Similar to recent studies [191, 211, 225], we set the significance level to $\alpha = 0.05$.

4.4.1 Performance on Individual Tasks

Breast Density Classification

We report an AUC score of 0.948 for density model B on the test set with TPR = 0.882 and TNR = 0.832 (F1 = 0.861). As depicted in Figure 4.9, the final breast density classification model B on the patient level (blue) shows a minor improvement in terms of AUC compared to the aggregated predictions mean(D) of the view model D on patient level with AUC = 0.943 ($p < 0.001$). mean(D) reaches a TPR of 0.833 and TNR of 0.889 (F1 = 0.858). Further, we observe a significantly higher sensitivity with D compared to mean(D) ($p < 0.001$) at similar accuracies as summarized in Table 4.3. On image level, we report an AUC of 0.924 with D (TPR = 0.815, TNR = 0.894, F1 = 0.849, accuracy = 0.854).

Table 4.3 also summarizes density classification results reported in the literature. We achieve higher accuracy scores on the DDSM dataset compared to Oliver et al. [168], who test their approach only on a subset of 831 R-MLO images, while our method is evaluated on 453 patients, i.e., 1812 view images. While our model performs slightly beneath published works, these methods are trained utilizing significantly larger datasets. For example, the NYU dataset by Wu et al. [251] comprises 200k exams with 80 % belonging to the training and 20 % to the test subset.

Findings Classification

For the task of classifying images into those with lesions and those without, model F reaches an AUC score of 0.921 on test data with TPR = 0.881 and TNR = 0.802 (F1 = 0.878).

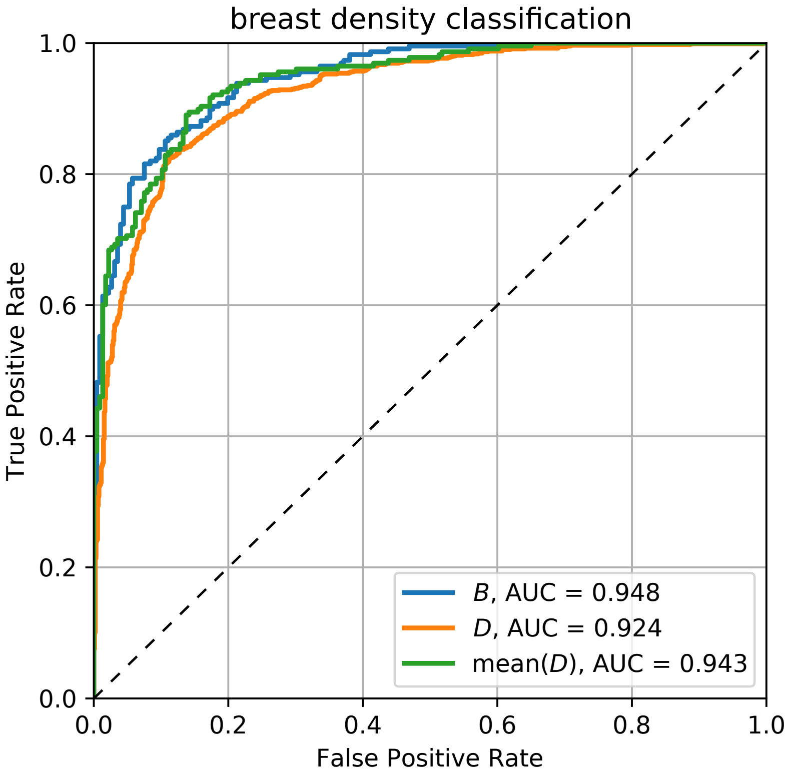


Figure 4.9: ROC curves comparing the breast density classification performance of the patient-level density model B, view model D, and mean(D).

Method	Data	Accuracy (4 classes)	Accuracy (2 classes)
Wu et al. [251]	private (NYU)	0.767	0.865 (derived)
Lehman et al. [123]*	private	0.770	0.870 (derived)
Kaiser et al. [102]	private	–	0.881
Oliver et al. [168]*	DDSM (R-MLO)	0.772	0.842
Ours* (D)	DDSM	–	0.854
Ours (mean(D))	DDSM	–	0.861
Ours (B)	DDSM	–	0.857

Table 4.3: Overview of reported density classification accuracies in related works and obtained with our model B. Methods indicated with * use one image as input, those without utilize all four view images.

Method	Train / Test Data	Lesion	TPR @ FPI
Agarwal et al. [5]	OMI-H / OMI-H	mass	0.93 @ 0.78
	OMI-H / INbreast	malignant mass	0.99 @ 1.17
		benign mass	0.85 @ 1.00
Ribli et al. [188]	DDSM, private / INbreast	malignant lesion	0.90 @ 0.30
Akselrod-B. et al. [9]	private / INBreast, private	mass	0.90 @ 0.30
Anitha et al. [14]	– / DDSM*	mass	0.925 @ 1.06
Ours (L)	DDSM / DDSM	malignant mass	0.84 @ 1.00
		malignant calcification	0.93 @ 1.09
		benign mass	0.70 @ 1.06
		benign calcification	0.68 @ 1.06

Table 4.4: Overview of lesion localization results reported in related works and results obtained with our model L (OMI-H = subset of images from OPTIMAM dataset acquired with a Hologic scanner, * = subset of 300 images used). The localization performance is given by TPR @ FPI.

To the best of our knowledge, there is only the work by Lotter et al. [144], who uses the presence/absence of lesions as classification target for pre-training their model on patch level. Thus, they do not report performance measures on image level.

Lesion Localization

We report TPR rates of 0.84 for malignant masses, 0.93 for malignant calcifications, 0.70 for benign masses, and 0.68 for benign calcifications by our localization model L on test images with lesions, as summarized in Table 4.4. Figure 4.10 shows the corresponding FROC curves. A lesion is considered detected if the intersection over union of the detected bounding box with the ground truth bounding box is ≥ 0.2, or, if the center of the detected bounding box lies within the ground truth bounding box [188]. On normal images in the test set (105 patients, i.e., 420 view images), we detect 386 false-positive lesions in 188/420 images. On the 348 abnormal cases, we detect 2478 false-positive lesions.

Figure 4.11 shows visual samples of correctly and falsely detected lesions. Overall, we report lower detection rates for benign lesions compared to malignant lesions, a phenomenon also observed in the literature [5]. As visible in Figure 4.11, one reason for the lower performance of model L is the detection of small calcifications that are highlighted in blue. They appear very similar to benign calcifications but are not annotated as such in the ground truth. Another aspect is the misclassification of denser breast tissue with masses as well as overlaps of benign and malignant masses that can occur due to non-maxima suppression performed on the class level.

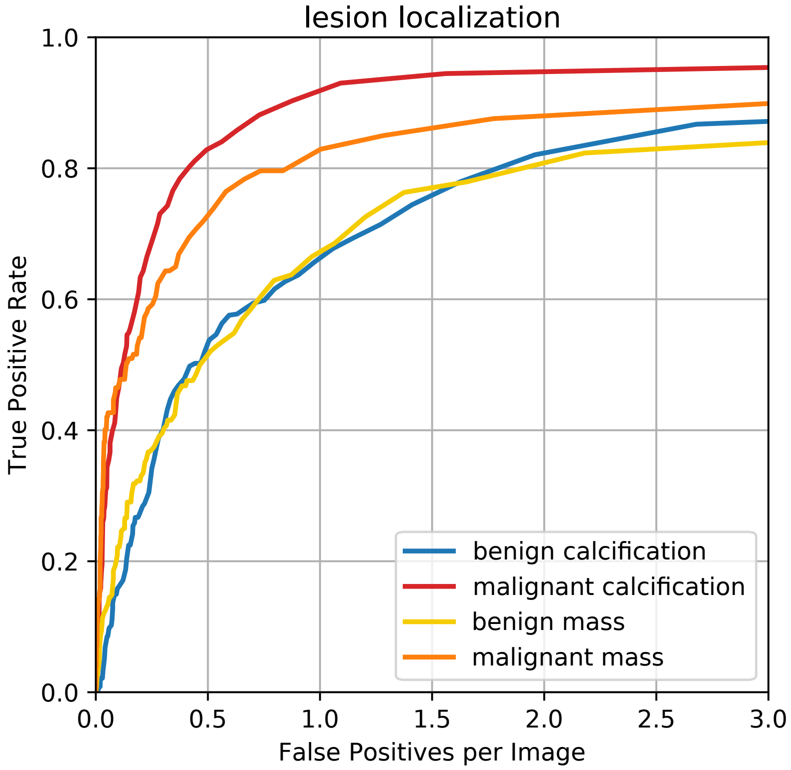


Figure 4.10: FROC curves comparing the performance of lesion localization model L for the four different classes.

Figure 4.12 shows another sample result for the right breast of a patient. This is a special case where the patient has a malignant calcification cluster but – according to the ground truth – it is visible in the MLO view only (ground truth = green). We investigate the performance of the lesion localization model L and the findings classifier F for this case. L detects benign and malignant calcifications in both breasts at the correct location, whereby the benign calcifications are incorrectly given a higher confidence compared to the malignant calcifications. According to the ground truth, nothing should be detected in R-CC. The findings model F wrongly classifies the R-CC view as image with a lesion, however, with a borderline score of $p_F^v = 0.503$, $v =$ R-CC (decision threshold = 0.5). Various patients are present in the dataset where lesions are annotated in only one view of a breast. To increase the reliability of lesion localization and image-wise classification models for such cases, symmetry aspects of the breasts could be considered, i.e., taking into account that the right and left CC view should have a similar appearance in case no lesions are present. However, precise ground truth data is necessary – especially in these special cases – to ensure correct predictions.

Table 4.4 provides an overview of localization results reported in the literature. We note that the localization performances of the different methods cannot be compared

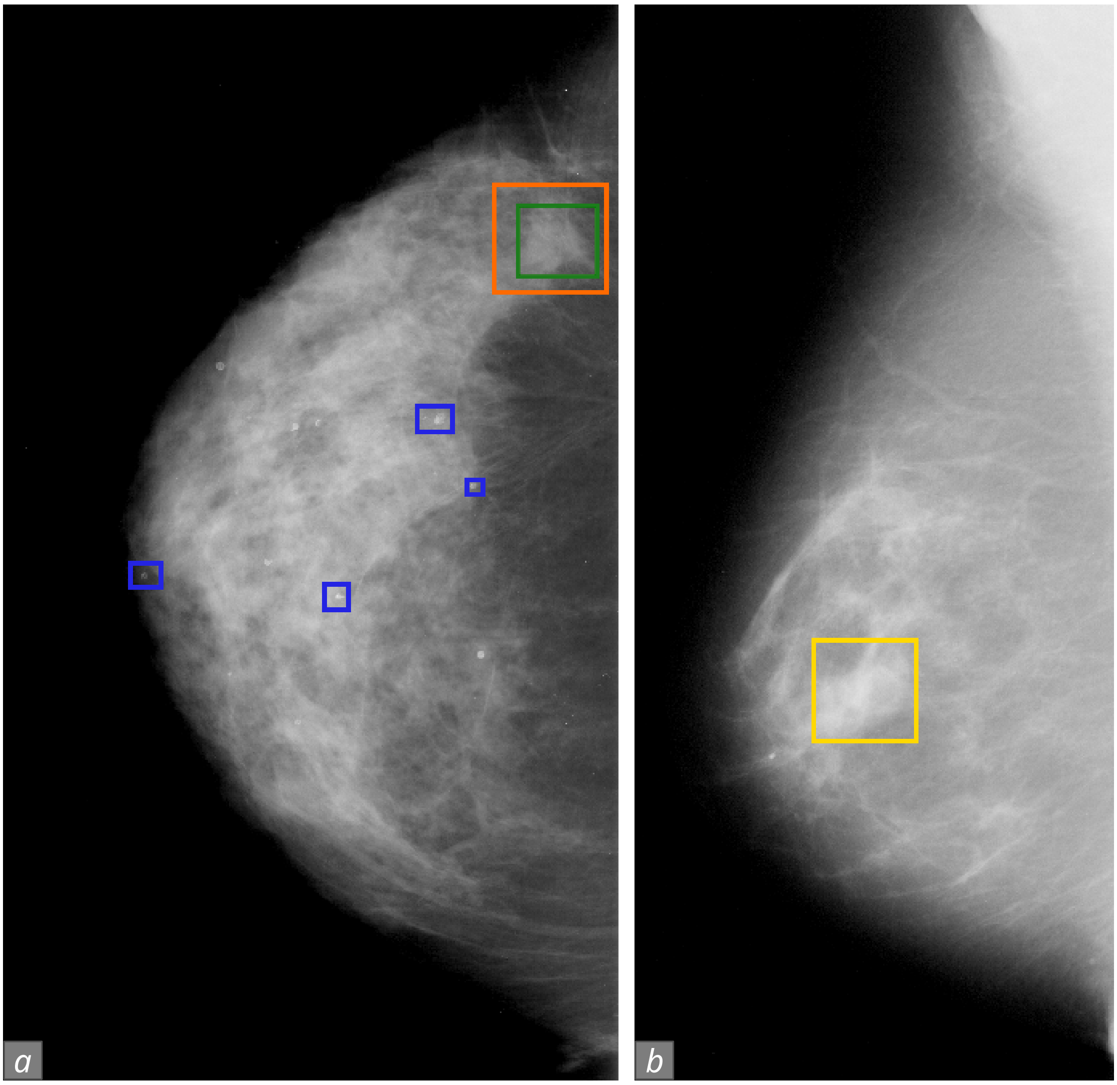


Figure 4.11: (a) R-CC image with correctly localized malignant mass (green = ground truth, orange = detected) and additional detected benign calcifications (blue) not present in the ground truth, (b) R-MLO image with false-positive benign mass (yellow).

directly due to the large differences in the datasets and varying criteria for correctly detected lesions. The method by Agarwal et al. [5], for example, utilizes the much larger OPTIMAM dataset, while the work by Anitha et al. [14], on the other hand, uses only a subset of the DDSM set rather than the full dataset. Despite these factors, we observe that overall our model L has a lower detection rate compared to the literature. Especially for benign lesions, our method falls behind the state-of-the-art.

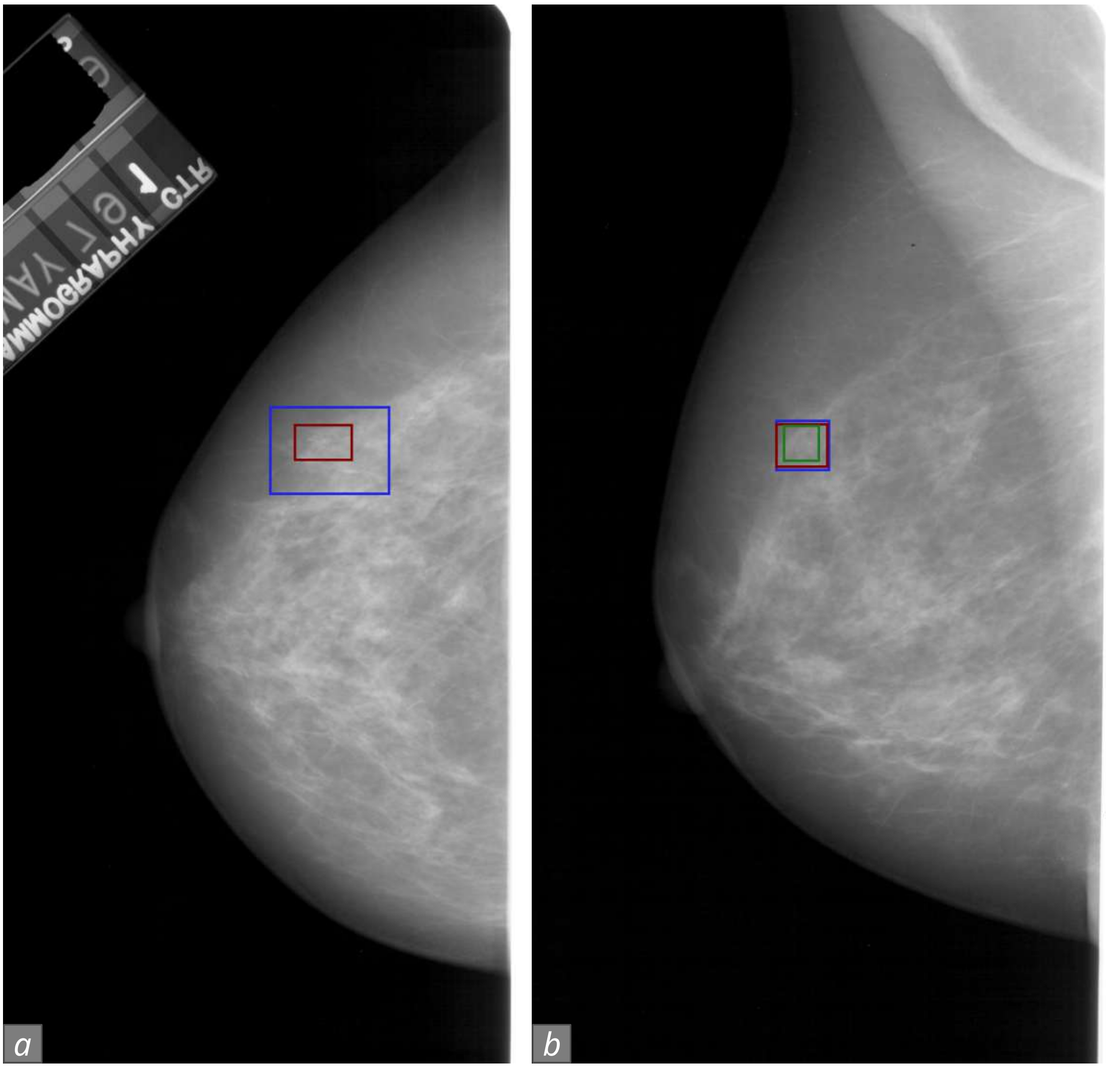


Figure 4.12: Localization result for the right breast of a patient where a malignant calcification in the form of a cluster (green = ground truth) is visible in the R-MLO view only. (a) R-CC image with false-positive localized benign calcification (blue) with high confidence of $p_L^{v,1} = 0.473$ and malignant calcification (dark red) with a lower confidence of $p_L^{v,2} = 0.148$, $v =$ R-CC, (b) R-MLO image with correctly localized malignant calcification (green = ground truth, dark red = detected) and additional false localization of benign calcification (blue). The benign calcification receives a higher confidence of $p_L^{v,1} = 0.261$ compared to the malignant prediction with $p_L^{v,2} = 0.180$, $v =$ R-MLO.

Target	Fusion	Model	AUC	F1	TPR	TNR
lesion	score	P_{score}	0.942	0.932	0.933	0.771
		P_{score}*	0.941	0.928	0.919	0.800
		$\max(p_F^v)$	0.922	0.938	0.974	0.667
	feature	P_{feat}	0.962	0.948	0.956	0.800
		P_{feat}*	0.959	0.943	0.939	0.829
		$\max(p_F^v)$	0.922	0.938	0.974	0.667
malignancy	score	P_{score}	0.778	0.601	0.591	0.813
		P_{score}*	0.774	0.523	0.578	0.857
		$\max(p_L^v)$	0.762	0.578	0.570	0.800
	feature	P_{feat}	0.791	0.603	0.638	0.763
		P_{feat}*	0.789	0.581	0.577	0.797
		$\max(p_L^v)$	0.762	0.578	0.570	0.800

Table 4.5: Performance metrics of patient fusion models P_{score} (MLPs) and P_{feat} on test data. Models marked with * indicate exclusion of breast density information. $\max(p_F^v)$ and $\max(p_L^v)$ denote the maximum of predictions scores of findings model F and localization model L for lesion prediction and malignancy prediction, respectively.

4.4.2 Patient Meta-Model Results

ROC curves for feature fusion with P_{feat} and score fusion with an MLP P_{score} for both patient predictions are shown in Figure 4.13. Table 4.5 summarizes quantitative performance measures on test data. We additionally train our fusion models without density information (indicated with * in Table 4.5) and compare all our fusion results to standard ensembling, i.e., taking the maximum of prediction scores. We perform paired statistical significance tests between all fusion models, including also standard ensembling. Table 4.6 and Table 4.7 summarize p-values, with a value of $p < 0.05$ denoting a statistically significant difference between two models.

For P_{score}, we obtain the best results in terms of AUC and TPR with MLPs for both patient predictions, compared to SVMs and Random Forests (see Table 4.8). In terms of the number of included lesions n in the meta-models, the best results reported in Table 4.5 and Table 4.8 are obtained with $n = 3$ for the lesion prediction (P_{score} and P_{feat}), and $n = 3$ (P_{score}) and $n = 1$ (P_{feat}) for the malignancy prediction. A detailed overview of quantitative results for P_{score} for SVM, Random Forest, and MLP, and P_{feat} for different numbers of included lesions n is provided in the Appendix.

Overall, we report an increase in terms of AUC between 0.02 and 0.04 for the lesion prediction with score fusion and feature fusion, respectively, when comparing to the score maximum across the four views $\max(p_F^v)$ ($p < 0.001$ for both). A slightly smaller increase is obtained for the malignancy prediction, ranging from 0.016 for P_{score} to 0.029 for P_{feat}, as compared to $\max(p_L^v)$ ($p < 0.001$ for both). We report higher AUC scores

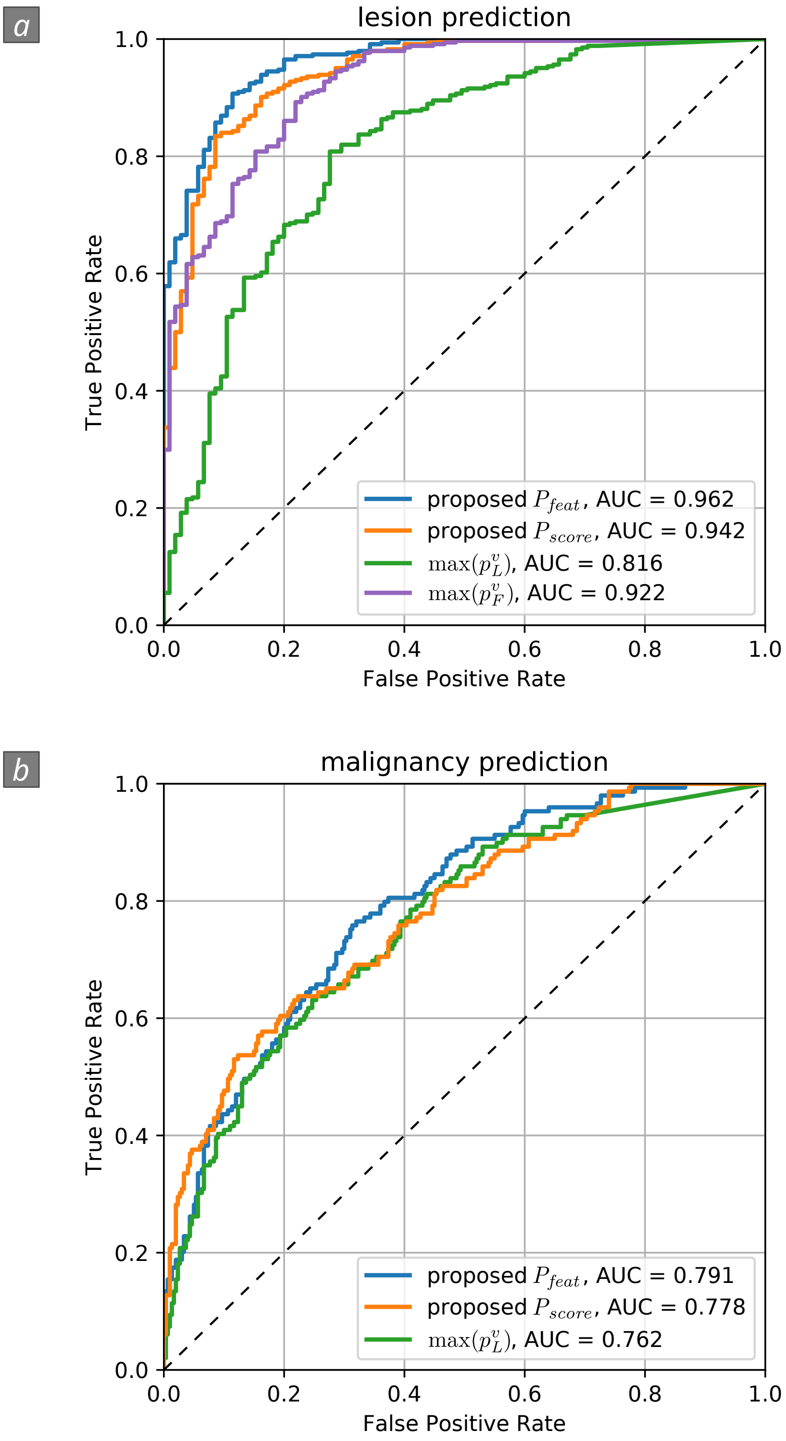


Figure 4.13: ROC curves of patient meta-models P_{feat} and P_{score} for the (a) lesion prediction, and (b) malignancy prediction. $\max(p_F^v)$ is excluded in (b) as we cannot derive malignancy information from F.

Model	P_{score}	P_{score}*	P_{feat}	P_{feat}*	$\max(p_F^v)$	$\max(p_L^v)$
P_{score}		**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
P_{score}*	**< 0.001**		**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
P_{feat}	**< 0.001**	**< 0.001**		**< 0.001**	**< 0.001**	**< 0.001**
P_{feat}*	**< 0.001**	**< 0.001**	**< 0.001**		**< 0.001**	**< 0.001**
$\max(p_F^v)$	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**		**< 0.001**
$\max(p_L^v)$	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	

Table 4.6: Statistical significance analysis for the lesion prediction: A p-value of $p < 0.05$ denotes a statistically significant difference between two respective models (bold font). Models marked with * indicate exclusion of breast density information. $\max(p_F^v)$ and $\max(p_L^v)$ denote the score maximum of findings model F and localization model L for the lesion prediction, respectively.

Model	P_{score}	P_{score}*	P_{feat}	P_{feat}*	$\max(p_L^v)$
P_{score}		**< 0.001**	**0.023**	**0.002**	**< 0.001**
P_{score}*	**< 0.001**		**< 0.001**	0.237	0.999
P_{feat}	**0.023**	**< 0.001**		**< 0.001**	**< 0.001**
P_{feat}*	**0.002**	0.237	**< 0.001**		0.920
$\max(p_L^v)$	**< 0.001**	0.999	**< 0.001**	0.920	

Table 4.7: Statistical significance analysis for the malignancy prediction: A p-value of $p < 0.05$ denotes a statistically significant difference between two respective models (bold font). Models marked with * indicate exclusion of breast density information. $\max(p_L^v)$ denotes the score maximum of localization model L for the malignancy prediction. $\max(p_F^v)$ is excluded in this analysis as we cannot derive malignancy information from F.

Target	Model	AUC	F1	TPR	TNR
lesion	MLP	0.942	0.932	0.933	0.771
	SVM	0.935	0.928	0.916	0.810
	Random Forest	0.929	0.924	0.919	0.771
malignancy	MLP	0.778	0.601	0.591	0.813
	SVM	0.763	0.552	0.483	0.867
	Random Forest	0.776	0.581	0.564	0.813

Table 4.8: Comparison of performance metrics of P_{score} for MLP, SVM, and Random Forests for both patient predictions.

and increased sensitivity, i.e., TPR, with feature fusion models compared to score fusion models for both patient predictions ($p < 0.001$). However, for the malignancy prediction, we observe a reduced specificity, i.e., TNR, for feature fusion as compared to score fusion.

Figure 4.14 and Figure 4.15 show sample results obtained with our proposed mammography pipeline. In the example in Figure 4.14, the localization model L is able to correctly localize the malignant mass in the right breast, but also falsely detects a benign mass at the same location. In the R-MLO view, the benign detection is given a higher confidence than the malignant detection, which would lead to false patient-level results in case we solely rely on L. However, the feature fusion models P_{feat} are able to correctly classify the patient in terms of the lesion and malignancy prediction, while P_{score} fails for the malignancy prediction (decision threshold $= 0.5$).

The second case (see Figure 4.15) shows a normal patient without lesions where our models perform incorrect predictions at different levels. The localization model L detects benign calcifications with a high confidence in the L-CC and L-MLO view, and a false-positive, low-confidence malignant mass in R-MLO. The predictions of the findings model F are in line with L, i.e., it falsely classifies the same three view images as images with lesions. Model F is most confident in the L-CC and L-MLO view where the localization model detects the high-confidence benign calcifications. This is probably the reason why the fusion models P_{feat} and P_{score} result in high-confidence, false predictions as well. For the malignancy prediction, P_{feat} is less confident in its decision compared to the models for the lesion prediction. With a p_P of 0.607 it is still above the decision threshold of 0.5, resulting in an incorrect prediction at the patient level. Only the score fusion model P_{score} delivers a correct assessment for the patient ($p_P = 0.268$).

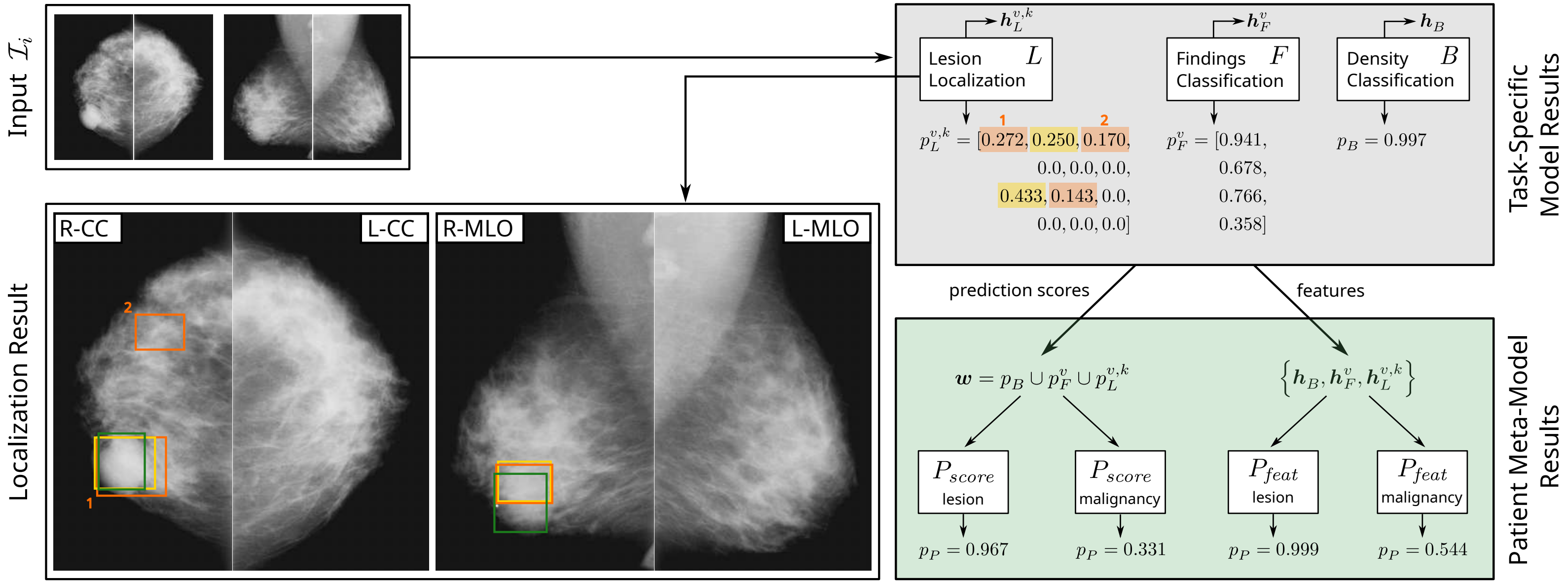

Input $\mathcal{I}_i$
Localization Result
R-CC
L-CC
R-MLO
L-MLO
1
2
Lesion Localization L
Findings Classification F
Density Classification B
$\boldsymbol{h}_L^{v,k}$
$\boldsymbol{h}_F^{v}$
$\boldsymbol{h}_B$
$p_L^{v,k} = [0.272, 0.250, 0.170, 0.0, 0.0, 0.0, 0.433, 0.143, 0.0, 0.0, 0.0, 0.0]$
$p_F^{v} = [0.941, 0.678, 0.766, 0.358]$
$p_B = 0.997$
Task-Specific Model Results
prediction scores
features
$\boldsymbol{w} = p_B \cup p_F^v \cup p_L^{v,k}$
$\{\boldsymbol{h}_B, \boldsymbol{h}_F^v, \boldsymbol{h}_L^{v,k}\}$
P_{score} lesion
P_{score} malignancy
P_{feat} lesion
P_{feat} malignancy
$p_P = 0.967$
$p_P = 0.331$
$p_P = 0.999$
$p_P = 0.544$
Patient Meta-Model Results

Figure 4.14: Sample result of our mammography pipeline: For an exam $\mathcal{I}_i$, we obtain task-specific model results first, i.e., the individual prediction scores p and the bounding boxes of localized lesions. Then, we feed prediction scores p and extracted features $\boldsymbol{h}$ to the dedicated patient meta-models P_{score} and P_{feat}, respectively, which results in the final patient-level prediction scores p_P. The illustration shows the benefits of our mammography pipeline: The patient has a malignant mass (green = ground truth) in the right breast. Model L is able to localize the malignant mass (orange), but with low confidence that is not sufficient to be reliably counted as detection. Low confidence localizations are also found by L for an additional malignant mass and a benign mass (yellow). Detection scores $p_L^{v,k}$ are colored according to their class label as detected by L. Results show that both fusion models P_{feat} are able to circumvent the low scores and correctly classify the patient ($p_P = 0.544$ for the malignancy prediction, $p_P = 0.999$ for the lesion prediction).

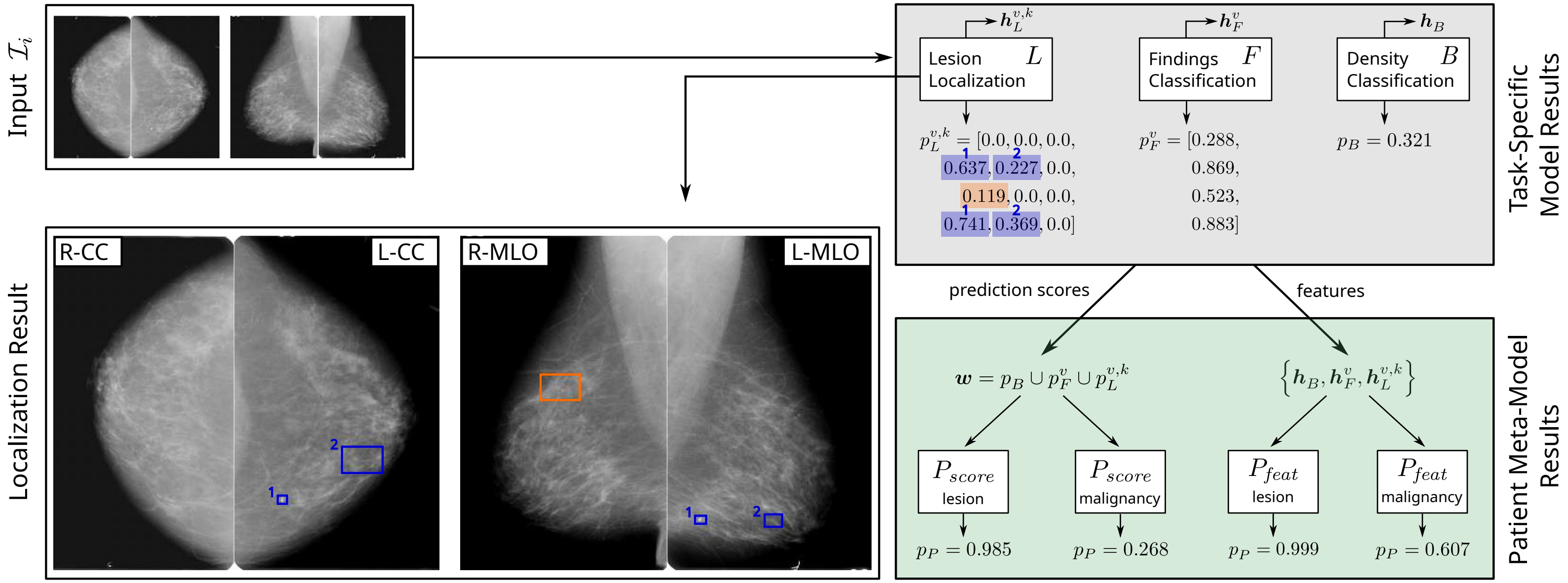


Figure 4.15: Sample result where our mammography pipeline delivers incorrect results on task and patient level for a normal patient: While the density model B correctly predicts non-dense breast tissue ($p_B = 0.321$), the localization model L falsely detects high-confidence benign calcifications (blue) in L-CC and L-MLO and a malignant mass (orange) with low confidence ($p_L^{v,1} = 0.119$ in $v =$ R-MLO). The detection scores $p_L^{v,k}$ are colored according to their class label. Further, the findings model F delivers incorrect predictions for the same three view images. Only the fusion model P_{score} for the malignancy prediction correctly classifies the full patient ($p_P = 0.268$) (decision threshold $= 0.5$).

4.4.3 Ablation Studies

Complementary to the training setup described in Section 4.3.1, we perform additional experiments to support our pre-training strategies for task-specific models B and F. Further, we investigate the influence of breast density information in the fusion models.

Pre-training of B

We retrain the density model B without pre-training the view model D with the same training parameters as summarized in Section 4.3.1, except for a lower learning rate of 1e-3. We obtain a significantly lower AUC score of 0.900 ($p < 0.001$) with this model, compared to the AUC of 0.948 achieved with B. Further we report TPR = 0.934, TNR = 0.690, F1 = 0.817, and a significantly lower accuracy of 0.797 as compared to B ($p < 0.001$).

Pre-training of F

Further, we retrain the findings classifier F without patch-wise pre-training with the same training parameters as described in Section 4.3.1. The model without pre-training achieves a significantly lower AUC score of 0.895 ($p < 0.001$) and sensitivity of 0.816, F1-score of 0.846, and specificity of 0.817.

Breast Density Ablation Study

As breast density is an essential risk factor for breast cancer [51], we investigate the effect of excluding this information form P_{feat} and P_{score}. We retrain our patient meta-models with the same training parameters, but exclude breast density features and scores for P_{feat} and P_{score}, and denote the obtained models $P_{feat}*$ and $P_{score}*$, respectively. Results in Table 4.5 show higher AUC scores and a higher TPR for all fusion models P_{score} and P_{feat} when *including* breast density information ($p < 0.001$, as summarized in Table 4.6 and Table 4.7). No statistically significant difference can be reported in comparing $P_{score}*$ and $P_{feat}*$ with $\max(p_L^v)$ for the malignancy prediction with p-values $p = 0.999$ and $p = 0.920$, respectively. Further, no significant difference can be observed between $P_{score}*$ and $P_{feat}*$ for the malignancy prediction ($p = 0.237$). These results indicate that the inclusion of breast density can yield improved classification performance.

4.4.4 Discussion

Breast Density

We investigate the patient density model B and aggregated view model mean(D) in depth and vary the decision threshold. The comparison of evaluation measures for both models is provided in Table 4.9. The results show that model B yields more reliable predictions with high confidence, and thus, higher sensitivities, accuracies, and F1 scores at various thresholds compared to the aggregated view model ($p < 0.001$). Such automated tools that deliver trustworthy, reproducible measures are of increasing

Threshold	Model	TPR	F1	TNR	Accuracy (2 classes)
0.6	B	0.860	0.871	0.885	0.872
	mean(D)	0.781	0.838	0.916	0.848
0.7	B	0.816	0.859	0.916	0.866
	mean(D)	0.711	0.804	0.942	0.826
0.8	B	0.763	0.841	0.947	0.855
	mean(D)	0.684	0.800	0.973	0.828

Table 4.9: Comparison of different measures obtained with B and mean(D) at various decision thresholds.

importance in clinical practice, especially for breast density assessment where subjectivity and high inter-observer variability are well-known issues [51, 102, 216]. As breast density is considered an important risk factor for the development of breast cancer, reliability and reproducibility are key aspects when it comes to standardized density reporting which may trigger supplemental/personalized screening procedures [45, 51, 216].

Comparison to Related Work

Table 4.10 sets our method in context to related approaches in the literature. In general, a direct comparison of reported evaluation measures of different methods is not possible as datasets used for training and evaluation differ vastly, e.g., with respect to varying imaging quality and modality (scanned film vs. Full-Field Digital Mammography (FFDM)), overall number of images, or amount of training data. To counteract this issue at least to some extent, we report train and test data in Table 4.10 and refer to the respective publications for further details. In addition, we compare our results to those reported with a single model or fusion model, and without test-time augmentation, i.e., results that are obtained by ensembling multiple predictions of the same model on augmented test data.

Fusion-Based Methods. Overall, our multi-input CNNs improve AUC scores by 0.029 to 0.040 compared to standard model ensembling. Similar increases for fusion approaches have also been reported in the literature. Kooi et al. [112] report an improved AUC by 0.019 when adding handcrafted features, like contrast or texture, to CNN features for the classification of single mammograms. The work by Kyono et al. [115] fuses multi-task scores, like “diagnosis”, “suspicion”, “conspicuity”, “breast density”, across multiple views, similar to our method. The difference is that their multi-task model predicts the same scores per view image, while we fuse predictions obtained from *different* models. Adding the multi-task output to their multi-view approach increases performance by 0.031 in terms of AUC. Shen et al. [208] fuse information on a single-image level in a weakly-supervised fashion, i.e., by fusing salient image regions with a fusion module, and report a single-model AUC score of 0.833 on CBIS-DDSM test data. A recent method by Lotter et al. [143] combines fully and weakly (multi-instance) supervised learning and reports state-

Method	Train / Test Data	Fusion Level	Result Level	Target	AUC	AUPRC
Shen et al. [206]	CBIS-DDSM / CBIS-DDSM	-	image	mal.	0.870	-
	CBIS-DDSM + INbreast / INbreast	-	image	mal.	0.950	-
Shu et al. [211]	CBIS-DDSM / CBIS-DDSM	-	image	mal.	0.838	-
	INbreast / INbreast	-	image	mal.	0.934	-
Ribli et al. [188]	DDSM, private / INbreast	-	breast (max+avg)	mal.	0.950	-
Lotter et al. [143]	DDSM, OPTIMAM, private / OPTIMAM	ROI	patient (max+avg)	mal.	0.963 ± 0.003	-
Kooi et al. [112]	private / private (NL screening)	ROI	image	mal. mass	0.941	-
Shen et al. [208]	CBIS-DDSM / CBIS-DDSM	image	breast (avg)	mal.	0.833	-
	private / private (NYU)	image	breast (avg)	mal.	0.891	0.390
Shachor et al. [205]	DDSM / DDSM	ROIs (CC+MLO)	breast	ben./mal. calc	0.661	-
Kyono et al. [115]	private / private (Tommy trial)	patient	patient	mal.	0.824 ± 0.016	0.580 ± 0.028
Ours (P_{feat})	DDSM / DDSM	patient	patient	mal.	0.791	0.660
		patient	patient	lesion	0.962	0.987

Table 4.10: Results obtained with fusion models P_{feat} compared to classification results reported in related works. Train and test data utilized by the respective methods are separated with "/".

of-the-art performance for mammogram classification (AUC = 0.963 ± 0.003, OPTIMAM data). To obtain a score on a patient level, they perform standard ensembling (average + maximum). Finally, McKinney et al. [154] average cancer risk scores that are predicted by an ensemble of three large-scale deep learning models (AUC = 0.889, OPTIMAM data). Each model fuses features at different stages and aggregates predictions in various ways, e.g., by considering the maximum score or via MLPs. In summary, the results in Table 4.10 show that our fusion model performs below related fusion-based methods in terms of the reported AUC scores for the malignancy prediction. We believe that improving the localization model will consequently result in a better performance on the patient level.

In terms of lesion prediction, we observe that – to the best of our knowledge – our method is the only one that specifically investigates this classification target. With an AUC score of 0.962 and F1-score of 0.948, this model could be reliably used, e.g., within a reporting system, where patients with lesions are examined first.

Non-Fusion-Based Methods. Apart from the summarized information fusion methods, there are numerous works that predict whether an image is malignant directly from a view image [206, 211], or/and additionally apply simple ensembling strategies for predictions on a breast or patient level [110, 188]. Shen et al. [206], for example, utilize patch-based pre-training and compare variants of ResNet and VGG in their work. They report an image-level AUC score of 0.87 on CBIS-DDSM and transfer-learned this model on the INbreast data where they reach an AUC of 0.95. Shu et al. [211] propose two region-based pooling strategies and achieve lower AUC scores on CBIS-DDSM (AUC = 0.838) and INbreast (AUC = 0.934) data as compared to Shen et al. [206]. Ribli et al. [188] localize suspicious lesions using Faster R-CNN and consider the maximum/average score on the image/breast level (AUC = 0.95 on INbreast data).

Fusion- vs. Non-Fusion-Based Methods. The results summarized in Table 4.10 show competitive performance of fusion- and non-fusion-based methods. Although there is no clear benefit of fusion-based approaches over non-fusion-based works in terms of AUC scores, fusion approaches show different advantages. Recent methods focus, for example, on the integration of radiological and clinical features or aim at increasing interpretability of models, which is an important aspect in the medical domain [21, 115, 143, 208]. These advantages, however, may come at the cost of more complex training procedures as compared to standard deep learning models [208]. One limitation of recent fusion-based approaches, including this work, is the requirement for detailed, high-quality expert-annotations [21, 112, 116, 154]. However, this is not limited to fusion methods per se, as the need for, e.g., bounding box annotations applies likewise to non-fusion-based, R-CNN/YOLO-based localization approaches [5, 9, 11, 188]. Recent weakly supervised works aim to tackle this issue and show already promising results [143, 208, 225].

Clinical Implications

In this chapter, we presented a technical proof-of-concept study for a mammography pipeline comprising of three task-specific models and patient meta-models that fuse task-specific features and predictions. While one goal is to obtain an improved assessment on a patient level as compared to standard model ensembling, the second goal is to develop a support tool for reading tasks of radiologists. Similar to recent technical proof-of-concept studies by Kyono et al. [115, 116] and Barnett et al. [21], we aim to provide intermediate results that are linked to radiological features and potential cancer risk factors. This is in contrast to studies that highlight the potential of workload reduction by excluding scans from reading that are very likely normal, i.e., do not have any suspicious lesions [65, 116, 169, 193]. Our global lesion and malignancy predictions could be used to prioritize images for reading instead of excluding them, and additional intermediate results of task-specific models can be presented to the clinicians during exam reading and diagnosis. Localizing suspicious lesions, for example, is an essential part when reading mammograms where clinicians examine both views and breasts [205]. Showing localized regions can aid radiologists in image interpretation [214], for example, by displaying only the most important findings and raising attention for them [191, 192]. Recent user studies in other domains, like prostate cancer diagnosis, confirm that prompting clinicians to suspicious regions helps them in reading [25]. In addition to the localized regions, our pipeline estimates the patient's breast density, which is an important risk factor for developing breast cancer [51], as mentioned earlier in this discussion.

Limitations

One limitation of this study is the relatively small DDSM dataset with only 2254 patients (after curation), as compared to resources available in related works [5, 115, 154, 208, 251]. Further, as the DDSM data consists of scanned film mammograms only, the imaging quality is significantly lower as compared to FFDM images. However, the usage of a fully open dataset fosters the development and comparability of approaches, while, e.g., access to the OPTIMAM dataset [74] and high-quality, expert-annotated data in general remains limited [169, 214]. A second factor may be the fixed number of detected lesions n currently used in our fusion models, which could be targeted with a multi-instance approach in the future. Finally, using a combined model that performs lesion and malignancy prediction in a multi-task fashion would reduce the number of models and could potentially also improve the performance of the malignancy prediction.

Future Perspectives

The transfer of the complete pipeline to a large FFDM dataset would be the logical next step and could potentially boost performance if trained on a larger data resource. To mitigate the requirement for expensive bounding box annotations for the lesion localization model L, an interpretable, weak localization approach similar to our recent works [125, 151] can be integrated in conjunction with the findings model F. Further, the improvement of the localization performance of model L, especially for benign lesions, as well as the training of one combined model for lesion and malignancy prediction are considered future work. The inclusion of additional radiological features, such as non-image-based risk factors, e.g., patient age, patient/family history, would be of interest and importance for clinical use [242]. Moreover, analyzing the temporal change of lesions is considered an important biomarker in practice [111, 201]. Finally, the evaluation of our proposed pipeline in a clinical reader study would help to evaluate further the potential benefits and risks of having global and local, task-specific information available, e.g., in terms of acceptance, increased interpretability, and potential bias [169].

4.5 Conclusion

In this chapter, we proposed the fusion of predictions and features from different *task-specific models* for improving mammography screening data classification. We train and evaluate the fusion models for two different classification targets relevant in the field of mammogram analysis: the prediction of *(i)* the presence of lesions and *(ii)* the presence of malignant lesions in a patient. Our experiments on public mammography data show that the fusion of scores with MLPs as well as feature fusion with multi-input embedding CNNs improves AUC scores compared to standard ensembling. Overall, we report an AUC score of 0.962 for predicting the presence of lesions and 0.791 for classifying the presence of malignant lesions on a patient level. By supporting our global predictions per patient with the local sub-results obtained by the task-specific models, we aim to aid clinicians in their reading and decision process. Finally, we perform an ablation study with breast density scores and features and conclude that additional density information can benefit the classification performance for both target scores.

CHAPTER 5

Concluding Remarks

5.1 Summary

In this thesis, we target the data heterogeneity characteristics of medical imaging data in image-based computer-aided detection and diagnosis applications with different machine learning-based methods. Inspired by literature from visualization research as well as from the medical image analysis domain [109, 148, 182], we first derived different *data heterogeneity categories* for medical imaging data in Chapter 1: multi-modal, -parametric, -dimensional, -resolution, -scale, -view, -temporal, -subject, -vendor or -center data [109, 148, 182]. We identified three general approaches how the different types of data from the various categories can be treated in the context of medical image analysis: (a) standard, (b) generalization-based, (c) and fusion-based approach. We discussed different concepts that follow strategies (b) and (c) and provided examples from the literature in medical image analysis, CADe, and CADx. In the core of this thesis, we presented different generalization- and fusion-based approaches in Chapter 2, Chapter 3, and Chapter 4 for two different application domains in the field of radiology: *anatomical labeling of the spine* and *mammography image analysis*. In the three chapters, we addressed a subset of the given data heterogeneity categories – specifically: multi-parametric, multi-modal, multi-vendor, multi-subject, multi-center, multi-view, and multi-scale data.

We dedicated Chapter 2 to thesis goals **G.1** and **G.2** and presented a method for anatomical spine labeling in multi-sequence, multi-vendor, multi-center MRI data. In contrast to related works, we aimed for a general solution that is applicable without retraining the method to various MR image contrasts – a problem that was not targeted in spine labeling research at that time. To achieve goal **G.2**, we utilize a model-based approach called ETMs. We propose to build local three-disc entropy models that are matched iteratively to the spine, starting from an initial, user-provided position. The intervertebral disc positions obtained from model matching are refined via an adaptive refinement method that is inspired by Haar-like features [233]. Our cross-validation evaluation setup demonstrates our pipeline's generalization capability and applicability to various multi-parametric MRI sequences covered and not covered by model training. One limitation of this solution is the semi-automated nature of the pipeline, as we require an initial click position and the corresponding anatomical label from a user.

In Chapter 3, we addressed this shortcoming and extended our approach towards a fully automatic solution. We utilize ETMs and CNNs to automatically detect the sacrum region in a scan as starting position for the labeling. Similar to Chapter 2, we perform an iterative matching of three-disc ETMs and a subsequent center position refinement step with an improved template matching approach. We extend the experiments compared to Chapter 2 and evaluate our pipeline approach on numerous public datasets. The data collections comprise not only multi-parametric MRI data, but also include multi-modal data, i.e., MRI *and* CT scans. Our results demonstrate the generalization capabilities of the proposed pipeline to new types of multi-parametric MRI sequences as well as to CT scans and vice versa. These results are directly related to thesis goal **G.3**, i.e., the applicability of methods to data from different modalities without retraining. One unsolved limitation is the dependence on the sacrum as initial position for the labeling. For a fully automated solution applicable to scans where the sacrum is absent, the detection of additional, reliable anchor positions is required.

Finally, we presented different fusion-based approaches in the context of multi-view screening mammography classification in Chapter 4, i.e., we address thesis goals **G.4** and **G.5**. The main objective is to improve patient-level classification by fusing information from different deep learning-based models, each dedicated to a specific mammography-related task and operating on a different scale of the data. This is in contrast to the standard model ensembling that is often performed. First, we introduce three task-specific models that classify breast density, the presence of lesions, and perform localization and classification of suspicious regions in a mammogram. Based on these task-specific models, we propose two fusion models that combine the individual model predictions and features on a patient level, respectively. Our experiments show that the fusion strategies improve AUC scores as compared to standard model ensembling.

5.2 Advances in Spine Labeling

In Chapter 2 and Chapter 3, we presented two methods for semi- and fully-automated semantic labeling of the spine that date back to 2016 and 2018. Since then, numerous articles have been published in this area. In the following, we review selected recent papers in Section 5.2.1 that focus on anatomical spine labeling in MRI data and address related challenges. For a comparison, we also mention methods presented for CT scans. An overview is provided in Table 5.1. In addition, we add two methods to the discussion that segment spinal structures and assign tissue class labels (last row in Table 5.1). Section 5.2.2 concludes this review and summarizes potential future research topics in spine labeling.

Finally, we refer the reader to recent surveys [23, 71, 90, 181], citing our work presented in

	Modality	
Task	MR	CT
Localization & Labeling	[10],[16]*,[227],[250],[266]	[39],[89],[95],[99],[134],[203]
Localization & Labeling & Segmentation	[56],[132],[170],[263]	[174],[224]
Segmentation & Labeling	[32],[50],[126],[127],[145], [231]	[126]
Segmentation & Tissue Class Labeling	[75],[114]	

Table 5.1: Overview of selected spine labeling literature (rows 1–3) and methods that segment spinal structures and assign tissue classes instead of anatomical labels (last row). The work marked with * cites our work presented in Chapter 3.

Chapter 3. They provide a broad overview of various topics in the field of spine imaging and image analysis with machine learning-based methods. These surveys also address related topics, such as the measurement of spine geometry and the grading of specific spinal diseases, which is not part of this discussion. One survey [71] covers related works from the complete musculoskeletal imaging workflow and discusses the potential impact of AI on it.

5.2.1 Discussion

We compare the methods summarized in Table 5.1 in the following aspects: the addressed task, applied learning-based method, targeted categories of heterogeneous data, approach to address data heterogeneity, assumptions of the methods, and the addressed research topics.

Task. According to a recent survey [23], the majority of recent methods in spine analysis focuses on segmentation tasks, which is also in line with our observations in this literature search. This development may be driven by recent vertebrae and disc segmentation and labeling competitions for MRI and CT data [202, 261, 267] and the availability of the respective annotated datasets. The benefit of semantic segmentation is that a labeled center position of the spinal structure can be derived if required. However, this is more resource-intensive as compared to pure localization approaches. Finally, the segmentation of vertebrae and discs oftentimes serves as basis for subsequent disease classification or grading, e.g., in the work of Lu et al. [145]. Regarding the tissue of interest, the presented papers usually focus either on discs or vertebrae. Only a few target both structures [127, 170] or segment multiple spinal tissues like discs, vertebrae, and spinal canal, but do not assign an anatomical label [75, 114].

Learning-Based Method. In the spine labeling methods presented in Chapter 2 and Chapter 3, we utilize a classical AAM-based approach and, in the latter, also a deep learning-based method, i.e., a CNN, in our pipeline. In contrast, the vast majority of recently developed algorithms for the analysis of the spine are purely deep learning-based, a trend that is also visible in recent vertebrae and disc segmentation and labeling competitions [202, 261]. From the methods summarized in Table 5.1, one uses regression forests [99], others combine deep learning approaches with classical methods, e.g., FCNs and Hidden Markov Models [39] or Faster R-CNNs with clustering approaches [89]. The general deep learning approaches which are utilized are, for example, region proposal networks [89, 231, 266], reinforcement learning [10, 263], graph convolutional networks [32], recurrent neural networks [75, 134, 227], or adversarial learning strategies [75, 203]. Further, FCNs are widely used, mainly for segmentation [50, 56, 114, 126, 127, 132, 145, 174, 227, 263], but have also been utilized for regression of, e.g., tissue centers [16, 39, 134, 174, 203, 250, 263]. The majority of FCNs are 2D/3D U-Net-based architectures. The method by Azad et al. [16], which cites our work, detects intervertebral candidate positions with a U-Net-based architecture combined with a shape-attention module. The authors use both the original image and the gradient image as model input and in this way incorporate shape information. In a second stage, they apply a false-positive reduction and labeling module inspired by YOLO [100]. Recently, also Transformer-based approaches have found their way into spine labeling research [224].

Multi-Modal, Multi-Parametric, Multi-Vendor, Multi-Center Data. In terms of the imaging modality, spine labeling methods are still usually developed in a modality-specific way. The applicability of models to the respective other modality was not investigated for the methods in Table 5.1. Only the method by Lessmann et al. [126] is trained and evaluated on CT and MR data, whereby modality-specific models are trained. In general, we observed that more diverse CT and MR datasets have been used more frequently in recent methods. While in 2016, Rak & Tönnies [184] criticized that MRI-specific challenges like changing image sequences and acquisition parameters are mostly not addressed explicitly, this has changed since then. Methods tailored to a certain MRI sequence are still being developed, e.g., for Dixon data [50, 56, 132] or for specific T2w datasets [126, 127, 170]. The use of multi-parametric, multi-vendor and often multi-center T1w and T2w scans has been more common recently [10, 32, 75, 114, 145, 227, 231, 250, 263, 266]. Methods for CT widely use data from recent competitions, like the MICCAI 2014 vertebrae localization [70] or the VerSe vertebra segmentation challenge dataset [135]. The latter comprises scans from different fields of view, acquired from multiple scanners, vendors, and centers.

Approach to Address Data Heterogeneity. In terms of generalization- and fusion-based approaches, we observe that the fusion of spine imaging data is usually performed for MR Dixon data, as demonstrated in several works [50, 56, 114, 132]. Recent disc segmentation challenges on that type of MR data are further driving the development of fusion approaches, as summarized, e.g., by Zeng et al. [261]. In Chapter 2 and

Chapter 3, we approach MR Dixon data differently and treat the four different image channels separately instead of fusing them. This allows us to use only 1/4 of the data as input at test time, which in practice is more memory efficient than methods that use all four channels as input to a multi-modal/multi-parametric CNN like the methods by Li et al. [131, 132], Dolz et al. [56], and Das et al. [50]. Two other approaches propose fusion-based methods for MR data. The works by Huang et al. [89] and Sekuboyina et al. [203] fuse coronal and sagittal images, whereby the former perform late fusion of detected vertebrae-center positions and the latter fuse features in intermediate layers with a butterfly-like FCN architecture.

The remaining methods use different approaches to address the heterogeneous nature of MRI data. Concepts like domain adaptation or continual learning that we discussed in Chapter 1 have not (yet) been adapted for spine labeling. Many approaches utilize data augmentation to increase the diversity of data samples [50, 126, 170, 227, 231, 250], while others do not perform any augmentation [56, 75] or do not mention anything in this regard in their work [10, 16, 32, 114, 132, 145, 263, 266]. Additional strategies to increase the generalization capabilities are the following: Vania & Lee [231] replace batch normalization layers in the ResNet-50 feature extractor with a group normalization and a dropout layer to handle the T1w and T2w data, i.e., to increase the generalization effect. Van Sonsbeek et al. [227] include gamma transforms in their data augmentation and argue that this encourages the network to learn the anatomy rather than intensity variations present in T1w and T2w data. Kuang et al. [114] utilize multi-scale feature learning and a feature distribution loss that forces the model to extract similar features for the same tissue in different MRI scans. Han et al. [75] use a modified autoencoder that enables them to deal with the high variability in the appearance of spinal tissue in MR data. Finally, in our works presented in Chapter 2 and Chapter 3, we also investigate the applicability of trained models to unseen datasets/domains as compared to the training data. This has been investigated only by two recent methods [114, 250].

Assumptions. The focus on a dedicated region of the spine is a common assumption in recent works, e.g., on the cervical [16] or lumbar and (lower) thoracic spine [10, 50, 56, 75, 114, 126, 127, 132, 145, 170, 231, 263]. Other assumptions include fixed numbers of present vertebrae or discs [16, 170, 263] or the presence of a certain vertebra like the sacrum [263]. Lu et al. [145] also have one dedicated network for sacrum segmentation, which is the basis for correct labeling of the spine. In the fully-automated work presented in Chapter 3, we also require a scan where the sacrum is present, but we are not restricted to lumbar scans. To counteract these limitations, more and more works are developed for arbitrary fields of view, avoiding the need for any assumption on the body region [32, 39, 89, 95, 99, 126, 134, 174, 203, 224, 227, 250, 266].

Research Topics. One central point of interest in recent spine labeling research is the problem of varying fields of view or incomplete scans. For CT, this has been a long-investigated problem that was also addressed in our research group in the work of Major et al. [149]. Also, in the last years, this has been a focus of research for CT data [39, 95, 99, 126, 134, 174, 203, 224]. Recently, this topic also gained momentum for spine labeling in MR scans [32, 250, 266], which was – to the best of our knowledge – not investigated when we were working on MR spine labeling. Due to oftentimes limited anatomical context, varying fields of view pose additional challenges. Chang et al. [32], for example, utilize global and local graph convolutional networks and a label attention network to reduce ambiguity of vertebrae. As mentioned before, the segmentation of the spine was also intensively investigated, whereby the focus was on the lumbar region. Baur et al. [23] conclude that lumbar spine segmentation was most prominent due to the high clinical relevance. Finally, the use of more diverse datasets can be considered a focus of recent MRI spine labeling publications.

5.2.2 Future Research Topics

The segmentation and anatomical labeling of spinal structures will remain an active field of research as it is an essential first step for the classification of pathologies and degenerative pathologies [23]. Furthermore, the reliable segmentation of *all* relevant anatomical structures in MRI is a potentially important future topic. Current models usually consider only discs or vertebrae, while models that segment and label multiple anatomical structures simultaneously are needed, like the works by Pang et al. [170] and Li et al. [127]. Additionally, the segmentation of multiple spinal structures, like spinal canal, neural foramen, vertebrae, discs, etc., must be considered, also in combination with anatomical labeling of intervertebral discs and vertebrae. The latter is missing in recent works [75, 114]. Regarding the body region, most research so far focuses on the lumbar and lower thoracic spine. This can be explained by the high clinical relevance and availability of public data, however, more research on the thoracic and cervical spine is definitely needed [23]. One open challenge also remains the correct anatomical labeling of the spine in fully automated approaches. Especially the correct detection of anchor vertebrae like $L5$ or $S1$ is subject to further research. One problem we encountered in this regard is the shift by one label (see Section 3.5.1), which is also reported, e.g., by Windsor et al. [250]. To further increase the generalization capabilities and, as a consequence, the tissue center localization and labeling accuracies, the use of highly heterogeneous datasets will continue in the future. The extension from multi-parametric data to multi-modal data could be further beneficial to increase labeling and localization performance [181].

From a clinical perspective, the development of annotation-efficient methods is another crucial factor. So far, the vast majority of approaches are fully supervised. From the related works in Table 5.1, only one method is unsupervised [114] and one weakly supervised [227]. Due to the availability of large, public datasets, the need for unsupervised, semi-, or weakly-supervised methods may be reduced. However, in clinical practice, strong labels, e.g., voxel-level segmentation masks, may not be available. The full annotation done by medical experts is time-consuming and very expensive. In case one wants to adapt methods to data from clinical practice, the level of supervision in the training is an important issue.

Task	Publications
Lesion Localization and/or Classification	[18] (R), [31] (SF,R,U), [85] (U), [129] (SF,U), [133] (U), [139] (SF,U), [147] (SF,U), [186] (U), [195]* (SF,U), [196]* (U), [257] (SF,U)
Cancer Detection and/or Classification (benign/malignant, BI-RADS, cancer yes/no, etc.)	[19] (SF,R,U), [29] (U), [137] (R), [120] (R), [226] (U), [229] (SF), [234] (R,P,U), [236] (P,R), [238] (U), [239] (SF,P), [240] (SF,U), [243] (SF), [259] (U)
Lesion/Cancer Segmentation	[137] (R), [129] (SF,U), [238] (U), [268]* (SF,U)
Breast Density Classification/Estimation	[73] (R), [226] (U)
Breast Cancer Risk	[223] (U), [235] (U), [256] (P)

Table 5.2: Overview of selected, recent mammography literature. The type of data which is used by the respective methods is provided in brackets: P = FFDM processed, R = FFDM raw, U = FFDM unclear, SF = scanned film. Publications marked with * cite our work presented in Chapter 4.

5.3 Recent Developments in Deep Learning-Based Analysis of Mammograms

The deep learning-based analysis of mammograms has been an active field of research since the publication of our work presented in Chapter 4. In Section 5.3.1, we give an overview of recent developments and trends as observed in selected papers published between 2021 – where the related work discussion in Section 4.1.1 stops – and 2023, when this thesis has been finalized. Table 5.2 summarizes the selected papers and the tasks they have been presented for. In Section 5.3.2, we discuss current and potential future research directions in CADe and CADx for mammography.

Further, we refer the reader to recent surveys, which focus on machine learning and deep learning-based methods in mammography in general [63], or on mammography and related X-ray imaging techniques, i.e., Digital Breast Tomosynthesis and CT [201, 232]. The survey by Luo et al. [146] cites our work [249] and gives an overview about ten years of research in breast cancer imaging. They focus not only on mammography but discuss applications for a wide range of data, like ultrasound, MR, or digital pathology images. Gastounioti et al. [64] review methods related to the topic of breast cancer risk. Finally, the work by Lamb et al. [118] summarizes commercial lesion detection and diagnosis applications for screening mammography.

5.3.1 Discussion of Recent Related Work

We give an overview on the methods summarized in Table 5.2 and discuss the following aspects: the learning-based methods and approaches the papers present, how data heterogeneity aspects are addressed, and what data has been used.

Learning-Based Methods and Approaches. The vast majority of methods presented in Table 5.2 is purely deep learning-based. However, some works combine CNNs or FCNs with classical methods. Like for example Maghsoudi et al. [73], who apply random forests and SVMs to classify superpixels obtained from a U-Net into fatty vs. dense for breast density estimation. Two papers that cite our work [195, 196] extract features from lesion patches, with common CNN feature extractors, perform feature selection/reduction, and then apply classical ML models, like SVMs, Naive Bayes, etc. for lesion classification.

All methods apply some form of supervised training, except for the study by Tan et al. [223] who investigate a weighting of scores obtained from already existing solutions in their study. Liu et al. [137] propose a weakly-supervised lesion segmentation method, trained with image-level labels, which builds on their recent work on weakly-supervised, interpretable mammogram classification [208]. Further, contrastive pre-training strategies are leveraged as well by some recent methods [29, 133, 259]. Various approaches adopt multi-task learning strategies to improve learning capabilities of their methods and to solve multiple tasks at once. Yang et al. [257], for example, incorporate biopsy information and BI-RADS scores for malignancy classification. Tardy & Mateus [226] include several classification tasks and reconstruct the input image to cope with poor and missing labels. You et al. [259] combine different classification tasks with contrastive pre-training tasks. Multi-task fusion models have been proposed as well, which combine segmentation and classification tasks [129, 229, 238, 268]. The recently published method by Zhong et al. [268] cites our work and is also similar to it. First, they train task-specific models for density classification, mass segmentation, and lesion classification. Second, the models are integrated in a multi-task fusion model to enhance the overall prediction.

Many different deep learning architectures and methods have been used for the different mammography tasks. Faster-RCNN or YOLO-based approaches have been applied by various works for lesion localization and classification [147, 186, 257]. Li et al. [133] utilize a fully convolutional one-stage object detection approach. Liu et al. [139] combine Mask-RCNNs and Graph Convolutional Networks for mass detection. GANs and adversarial training strategies have been applied, for example, for unsupervised domain adaptation [239] or to ensure consistent predictions across different vendors [256]. Li et al. [133] use CycleGANs to generate images in the style of different vendors for pre-training. Another method [120] uses conditional GANs to generate corresponding contralateral images, i.e., images acquired from the same view but from the corresponding other breast, to make use of differences in the breast tissue for cancer detection. Autoencoders have been utilized to reliably generate images without lesions in an anomaly detection setup [85] or in a multi-task learning scenario as image reconstruction component [225].

Walsh & Tardy [234] use autoencoders for synthesizing abnormalities such as masses or calcifications into benign images. Besides all different kinds of CNNs, like ResNet or DenseNet, etc., prototype networks are used as well in combination with an EfficientNet CNN for interpretable cancer classification [236]. U-Nets are applied for different segmentation tasks, e.g., for pectoral muscle segmentation [73] or mass or cancer segmentation [229, 238]. Transformer-based architectures have found their way into the field of mammogram analysis as well. They are applied, for example, as encoder modules in segmentation tasks [268] or to aggregate image encodings extracted from all four views [256], i.e., L-CC, R-CC, L-MLO, and R-MLO.

Approach to Address Data Heterogeneity. Both fusion- and generalization-based approaches are widely used due to the multi-view nature of mammography data, whereby strategies related to fusion are particularly popular. This was already the case when we were working on the analysis of mammograms, as reviewed in Section 4.1.1. One popular strategy is the intermediate fusion where different input branches relate to different views. This is often performed for full exams [223, 235, 256, 259] or subsets of views, like the fusion of CC and MLO view of one breast [133, 147, 186, 239]. Some recent methods also fuse different views, e.g., a main view with one or more auxiliary views [29, 120, 139, 257]. This is done to integrate prior medical knowledge in a model, e.g., to exploit symmetry properties. For example, the L-CC and R-CC view should have a similar appearance in healthy patients. Yang et al. [257], for example, combine different strategies to improve mass detection and malignancy classification. In one part of their network, they encode CC-images separately and then fuse them, in the second part they encode images from the same breast separately and fuse them, and finally, fuse both subnetworks.

Late fusion, i.e., score fusion, is performed for example for breast density estimation [73] or cancer risk scoring [223]. The method by Tan et al. [223] fuses multi-modal data, i.e., 3D ultrasound and mammography cases, for cancer risk scoring. Already existing systems are applied and the scores from both systems are weighted. Similarly, Wanders et al. [235] train a small MLP where they combine different scores, generated by a commercially available and an open-access system. This is similar to our prediction score fusion model presented in Chapter 4.

Other ways to fuse information are, e.g., for multi-temporal or multi-scale data. The work by Bai et al. [19] addresses fusion of temporal data where one branch relates to a view image of the current year and the second one to the corresponding image of the previous year. Li et al. [129] feed two lesion patches at different scales to two different branches, where one performs segmentation and the second one classification of masses. Both branches are fused for prediction of the final diagnosis.

Methods that take only one image or a ROI patch as input often perform fusion as well, for example, feature fusion at various levels in the network [226], or the fusion of features from different stacked ensembles [18]. Some methods also fuse image information with intermediately obtained segmentation information [229, 268] or with saliency maps [137]. Another way to utilize knowledge from different tasks without fusion is for example via knowledge distillation to learn an interpretable prototype network [236]. Further, multi-task learning approaches can leverage information from different branches. The multi-task learning approach by Wang et al. [238] performs simultaneous breast tumor classification and segmentation, whereby the classification branch supports learning of the tumor segmentation task at different locations in the network.

Generalization-based approaches have also been investigated recently. For models that take only one mammogram as input, generalization across the four standard view images, i.e., L-CC, R-CC, L-MLO, R-MLO, is obtained by training jointly on all views instead of training a view-specific model. This is the standard approach, which is done by various methods, e.g., Walsh & Tardy [234], Wang et al. [238], among many others.

To increase generalization capabilities, Bai et al. [19] perform a pre-training on a mix of public datasets and execute the downstream task of breast cancer detection and classification on a private in-house dataset. Castro et al. [31] investigate different types of symmetry-based regularization, e.g., invariance in the loss function or equivariant model architectures. A few works are concerned with domain adaptation and generalization [133, 239, 256]. Li et al. [133], for example, utilize a multi-style contrastive learning strategy where they use CycleGANs to generate different styles of a mammography view image that represent images from different vendors. Finally, several works demonstrate the applicability to data distributions unseen during training by testing the presented methods on unseen mammography datasets [31, 85, 234, 236].

Data. We observed that public mammography image datasets like DDSM [77, 78], CBIS-DDSM [121, 122], and INbreast [161] are still widely used in recent works, although the number of exams and images is limited, compared to larger collections such as OPTIMAM [74]. Another drawback of CBIS-DDSM and DDSM is the lower image quality as they consist of scanned film mammograms. On the other hand, a license agreement is required for OPTIMAM [140], thus, restricting and limiting the possibility to use them for research purposes.

Fortunately, new multi-center, multi-vendor FFDM datasets have been published in the last years [26, 61, 163], which aim to diversify the landscape of publicly available data. The provided high-quality imaging data and annotations should foster the development of a wide range of different applications. In 2022, the *VinDr-Mammo* dataset [163] has been introduced. It comprises 5000 “for presentation”, i.e., processed for display, exams that were acquired in Vietnam. Lesion-level annotations are available for non-benign lesions, e.g., BI-RADS scores and bounding boxes, as well as breast-level assessments, such as breast density. The *CMMD* dataset [26] – the Chinese Mammography Database – comprises 3712 raw mammograms from 1775 patients and is split into two subsets. The first subset includes biopsy-proven benign and malignant tumors along with other clinical data. The second subset contains only malignant tumors along with molecular subtypes. The *ADMANI* dataset [61] is a large-scale, longitudinal dataset, acquired in Australia. It consists of more than 4.4M “for presentation” processed images from 629,863 patients which are split into three subsets. Apart from the imaging data, detailed patient data like demographics as well as histopathology data are available. A major strength of this dataset is the provided ground truth. Patients with indication of cancer were recalled for further assessment where the cancer was histopathologically confirmed or falsified. Patients with no signs of cancer were invited for routine rescreening after two years. This protocol results in a strong ground truth as false positive or false negative assessments as well as true positive and true negative cases can be identified. This enables the use of this dataset for a diverse set of tasks, including interval cancer detection, or for retrospective and prospective studies. The dataset is planned to be made publicly available, whereby a subset has been used already in a recent breast cancer detection competition.

5.3.2 Current and Future Research Directions

As reviewed in Section 5.3.1, various recent methods focus on fusion-related research topics, i.e., what to fuse and where to perform the fusion. Several methods investigate how to incorporate medical domain knowledge, an area of research that has become increasingly popular in the last years. One aspect is how to exploit symmetry that should be present between the two CC or MLO views of the breast, another one is how to utilize the complementarity between the CC and the MLO view. To achieve this, various ways how to model the relation between different views are investigated [120, 139, 147, 186, 257]. Further, the integration of prior knowledge about lesions is a current and promising future topic of interest. For example, the guided synthesis of lesions [234] or the integration of boundary properties of lesions that are obtained, e.g., through segmentation, are explored [129]. Recently, several methods [129, 229, 238, 268] incorporate segmentation of lesions in their approaches. The reported results show promising directions, but large-scale evaluations are needed to better understand the impact of these methods.

In terms of data, several efforts have been made to increase the size and diversity of publicly available datasets, as reviewed in Section 5.3.1. The public availability of validation datasets to compare different methods is another important future aspect as well as evaluating the reliability of methods in case of out-of-distribution data or adversarial attacks [232]. Further, the robustness of deep learning methods against variations in mammographic images, e.g., due to different vendors and proprietary processing, is essential [64]. Another topic to mention is the high-resolution characteristic of mammograms and how to deal with it. A few methods address this topic [137, 234, 243, 257], e.g., one work investigates how to better exploit pre-trained features in the context of high-resolution mammogram classification [243]. Since abnormalities in mammograms can be very subtle, e.g., calcifications and clusters thereof, taking the high-resolution nature of the data into account remains crucial. The importance of multi-modal approaches is highlighted as well in recent works [6, 146]. Ultrasound is often used as a complementary modality in diagnosis. Hence, the fusion of mammography and ultrasound data could be a promising next step towards improving the performance of deep learning-based breast cancer screening methods. Not only the fusion of imaging data, but also the combination of imaging data with clinical information, reports, or biomarkers can potentially improve the performance and should be considered in future research [146].

All methods in Table 5.2 are retrospective studies. In terms of clinical adoption, prospective studies are needed to evaluate the efficacy of current CADe and CADx methods in clinical practice [232]. In this regard, it is also important *how* AI-based systems are integrated into the workflow. This is in general a non-trivial question as it can also influence a reader. Yoon & Kim. [258] summarize various methods how the integration could be performed for different scenarios, e.g., single reading, double reading, and stand-alone reading, i.e., workload triage.

5.4 Outlook

In the field of computer-aided detection and diagnosis, the possibilities to advance existing methods and develop new approaches specific to a certain anatomy, bodypart, or disease are almost endless. In the following, we summarize a few potential future research directions we identify based on the publications presented in this thesis.

Clinical Applications. In Section 2.4, Section 3.6, and Section 4.4.4, we discussed potential future directions for the two clinical applications we targeted. For spine labeling, we already addressed one issue in Chapter 3, namely the advance from semi-automatic labeling towards a fully automated approach. One unsolved limitation is the dependence on the sacrum as initial object for the labeling. For a fully automated solution, the detection of additional, reliable anchor positions is required. Various open questions in spine labeling remain, especially in MRI, like the labeling of arbitrary field of view scans, as briefly discussed in Section 5.2. In mammography image analysis, the transition/generalization of our proposed method to full-field digital mammography, which exhibits better image quality and contrast, would be the logical next step. Further, the investigation of automated methods for 3D Digital Breast Tomosynthesis data is of high interest, also from a clinical perspective. Recent studies show that combining breast tomosynthesis and full-field digital mammography data reduces the recall rate for follow-up exams and can improve cancer detection rates [152].

Data Heterogeneity. Regarding the data heterogeneity and approaches how to address it, we will focus future research on the robustness and generalization capabilities of methods. In Chapter 2 and Chapter 3, we investigated the use of a model-based approach, namely ETMs, to achieve generalization across multi-parametric and multi-modal data. In the future, we plan to exploit deep learning-based, model-agnostic approaches such as continuous learning to address the challenges these heterogeneity categories bring to method development. We want to focus on the domain-shift problem in the data domain that we are facing in clinical environments. There we are confronted with changing acquisition protocols, varying scanning parameters, and scanners from different hospitals. One vision in this regard is to develop strategies how a continuous update of already trained models can be handled and integrated in a practical clinical environment.

State-of-the-Art Reports. In Chapter 1, we described the heterogeneous nature of medical imaging data and identified three general approaches to address it in the context of medical image analysis. However, we only discussed the main concepts for the generalization and fusion approach and selected studies from the literature. A more in-depth analysis of the problem space and related literature based on the structure we derived would be of interest. This could be potentially published as a state-of-the-art report or book chapter. Further, we summarized and discussed recent trends in spine labeling and the analysis of mammograms in Section 5.2 and Section 5.3, respectively. Both systematic analyses could be extended to more comprehensive surveys.

Clinical Studies. In the work presented in this thesis, we focused on technical proof-of-concept studies. The integration of our methods into the radiology image reading workflow is pending, which would be required to perform clinical studies that test the applicability of our methods in practice. This includes studies with radiologists and clinicians using our methods in their daily work where potential clinical benefits can be evaluated, like increased time efficiency or increased confidence in diagnosis. Apart from that, studies from a human-computer interaction perspective are required that evaluate *how* the approaches should be integrated to yield the best results. In the literature, two general strategies are described how the integration can happen, namely “expert first” and “computer first” [180]. In an “expert first” approach, the computer result becomes available after the initial diagnosis by the clinician, who eventually refines the decision afterwards. In contrast, in the “computer first” approach, the result of the CADe/CADx system is already available at the time of diagnosis. However, this could influence the reporting physician and potentially lead to biases that must be considered [180, 232].

List of Figures

List of Tables

Acronyms

AAM Active Appearance Model. 8, 19, 29, 67, 110

AI Artificial Intelligence. 69, 70, 109, 118

AUC Area Under Curve. 21, 87–89, 95, 97, 101–104, 106, 108, 161–163

AUPRC Area Under the Precision-Recall Curve. 89, 103

CADe Computer-Aided Detection. 2, 5, 7, 8, 107, 113, 118, 119

CADx Computer-Aided Diagnosis. 2, 5, 7, 8, 107, 113, 118, 119

CBIS-DDSM Curated Breast Imaging Subset of Digital Database for Screening Mammography. 4, 21, 74, 75, 77–79, 102–104, 121

CC Craniocaudal. 4, 70, 71, 73, 77, 92–94, 98, 100, 103, 115–117, 121, 124, 125

CNN Convolutional Neural Network. 14–17, 20, 43, 46, 47, 53–55, 70, 72–76, 102, 106, 108, 110, 111, 114, 115, 121, 161–163

CT Computed Tomography. 1–3, 5, 9, 10, 14–16, 20, 22, 27, 41–44, 51, 52, 55, 56, 63–66, 108–111, 113, 121, 123, 124, 127

DDSM Digital Database for Screening Mammography. 21, 74, 75, 77–79, 89–91, 93, 103, 105

DICOM Digital Imaging and Communications in Medicine. 43, 65

ETM Entropy-Optimized Texture Model. ix, xi, 8, 19, 20, 28–32, 35, 37, 44, 46–49, 51, 53–56, 66, 107, 108, 118, 122

FCN Fully Convolutional Network. 14, 43, 65, 110, 111, 114

FFDM Full-Field Digital Mammography. 74, 102, 105, 113, 116, 127, 128

FPI False Positives per Image. 89, 91, 127

FSE Fast Spin Echo. 42, 52

GAN Generative Adversarial Network. 16, 114, 116

INR Implicit Neural Representation. 12, 15, 16

MC Microcalcification. 70, 72, 124

MLO Mediolateral oblique. 4, 70, 71, 73, 77, 89, 90, 92–94, 98, 100, 103, 115–117, 121, 124, 125

MLP Multilayer Perceptron. 16, 75, 76, 88, 95, 97, 104, 106, 115, 128, 161–163

MR Magnetic Resonance. 9, 10, 16, 19, 20, 27, 28, 30, 31, 33–37, 39, 41–44, 51, 52, 56, 58, 64–67, 107, 110–113, 122

MRI Magnetic Resonance Imaging. xi, 1–5, 8–10, 12, 14, 16–20, 22, 27, 42, 59–62, 107–112, 118, 121, 123, 127

MRT Magnetresonanztomographie. ix

NeRF Neural Radiance Field. 12, 15, 16

PDF Probability Density Function. 29, 30

PDw Proton Density weighted. 42, 52, 65

PET Positron Emission Tomography. 1–3, 5, 14, 22, 127

RBF Radial Basis Function. 53, 54, 88, 161–163

ROC Receiver Operating Characteristic. 89, 90, 95, 96, 124, 125

ROI Region of Interest. 3, 15, 73, 74, 76, 82–84, 103, 115, 124

SVM Support Vector Machine. 53, 55, 74, 88, 95, 97, 114, 128, 161–163

SWA Stochastic Weight Averaging. 86, 87

T1w T1-weighted. 3, 9, 10, 12, 19, 20, 27, 28, 30, 34, 36, 37, 39, 42, 43, 51, 52, 59, 64–66, 110, 111, 121–123

T2w T2-weighted. 3, 9, 10, 12, 19, 20, 22, 27, 28, 30, 34–37, 39, 42, 43, 51, 52, 61, 64–66, 110, 111, 121–123

TIRM Turbo Inversion Recovery Magnitude. 27, 39, 42, 52

TNR True Negative Rate. 88, 89, 95, 97, 98, 101, 102, 162, 163

TPR True Positive Rate. 88, 89, 91, 95, 97, 98, 101, 102, 127, 162, 163

TSE Turbo Spin Echo. 52

Bibliography

[1] Challenge (Automatic Intervertebral Disc Localization and Segmentation from 3D T2 MRI Data). In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2015)*, volume 9402 of *Lecture Notes in Computer Science*, pages 107–158. Springer International Publishing, 2016.

[2] DDSM Tools. https://github.com/fjeg/ddsm_tools, Last update: 7 October 2015. Accessed: 25 May 2023.

[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/, 2015. Software available from tensorflow.org. Accessed: 31 May 2023.

[4] Richa Agarwal, Oliver Díaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic Mass Detection in Mammograms Using Deep Convolutional Neural Networks. *Journal of Medical Imaging*, 6(3):031409, 2019.

[5] Richa Agarwal, Oliver Díaz, Moi Hoon Yap, Xavier Lladó, and Robert Martí. Deep Learning for Mass Detection in Full Field Digital Mammograms. *Computers in Biology and Medicine*, 121:103774, 2020.

[6] Richa Agarwal, Moi Hoon Yap, Md. Kamrul Hasan, Reyer Zwiggelaar, and Robert Martí. Deep Learning in Mammography Breast Cancer Detection. In Niklas Lidströmer and Hutan Ashrafian, editors, *Artificial Intelligence in Medicine*, pages 1287–1300. Springer International Publishing, 2022.

[7] Agfa-Gevaert Group. Agfa HealthCare. http://www.agfahealthcare.com. Accessed: 2 April 2023.

[8] Anne M. R. Agur and Arthur F. Dalley. *Grant's Atlas of Anatomy*. Lippincott Williams & Wilkins, 13th edition, 2012.

[9] Ayelet Akselrod-Ballin, Leonid. Karlinsky, Alon Hazan, Ran Bakalo, Ami Ben Horesh, Yoel Shoshan, and Ella Barkan. Deep Learning for Automatic Detection of Abnormal Findings in Breast Mammography. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS 2017)*, volume 10553 of *Lecture Notes in Computer Science*, pages 321–329. Springer International Publishing, 2017.

[10] Walid Abdullah Al and Il Dong Yun. Partial Policy-Based Reinforcement Learning for Anatomical Landmark Localization in 3D Medical Images. *IEEE Transactions on Medical Imaging*, 39(4):1245–1255, 2020.

[11] Mohammed A. Al-Masni, Mugahed A. Al-Antari, Jeong-Min Park, Geon Gi, Tae-Yeon Kim, Patricio Rivera, Edwin Valarezo, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Simultaneous Detection and Classification of Breast Masses in Digital Mammograms via a Deep Learning YOLO-Based CAD System. *Computer Methods and Programs in Biomedicine*, 157:85–94, 2018.

[12] Raja' S. Alomari, Jason J. Corso, and Vipin Chaudhary. Labeling of Lumbar Discs Using Both Pixel- and Object-Level Features With a Two-Level Probabilistic Model. *IEEE Transactions on Medical Imaging*, 30(1):1–10, 2011.

[13] Raja' S. Alomari, Subarna Ghosh, Jaehan Koh, and Vipin Chaudhary. Vertebral Column Localization, Labeling, and Segmentation. In *Spinal Imaging and Image Analysis*, volume 18 of *Lecture Notes in Computational Vision and Biomechanics*, pages 193–229. Springer, 2015.

[14] J. Anitha, J. Dinesh Peter, and S. Immanuel Alex Pandian. A Dual Stage Adaptive Thresholding (DuSAT) for Automatic Mass Detection in Mammograms. *Computer Methods and Programs in Biomedicine*, 138:93–104, 2017.

[15] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, and Miguel Angel Guevara Lopez. Representation Learning for Mammography Mass Lesion Classification With Convolutional Neural Networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, 2016.

[16] Reza Azad, Moein Heidari, Julien Cohen-Adad, Ehsan Adeli, and Dorit Merhof. Intervertebral Disc Labeling With Learning Shape Information, A Look Once Approach. In *Predictive Intelligence in Medicine (PRIME 2022)*, volume 13564 of *Lecture Notes in Computer Science*. Springer, 2022.

[17] Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H. Gandomi. A Review on Multimodal Medical Image Fusion: Compendious Analysis of

Medical Modalities, Multimodal Databases, Fusion Techniques and Quality Metrics. *Computers in Biology and Medicine*, 144:105253, 2022.

[18] Asma Baccouche, Begonya Garcia-Zapirain, and Adel S. Elmaghraby. An Integrated Framework for Breast Mass Classification and Diagnosis Using Stacked Ensemble of Residual Neural Networks. *Scientific Reports*, 12:12259, 2022.

[19] Jun Bai, Annie Jin, Tianyu Wang, Clifford Yang, and Sheida Nabavi. Feature Fusion Siamese Network for Breast Cancer Detection Comparing Current and Prior Mammograms. *Medical Physics*, 49(6):3654–3669, 2022.

[20] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[21] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. A Case-Based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

[22] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, 58:82–115, 2020.

[23] David Baur, Katharina Kroboth, Christoph-Eckhard Heyde, and Anna Voelker. Convolutional Neural Networks in Spinal Magnetic Resonance Imaging: A Systematic Review. *World Neurosurgery*, 166:60–70, 2022.

[24] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de La Iglesia-Vayá. PadChest: A Large Chest X-Ray Image Dataset With Multi-Label Annotated Reports. *Medical Image Analysis*, 66:101797, 2020.

[25] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 'Hello AI': Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):104, 2019.

[26] Hongmin Cai, Jinhua Wang, Tingting Dan, Jiao Li, Zhihao Fan, Weiting Yi, Chunyan Cui, Xinhua Jiang, and Li Li. An Online Mammography Database with Biopsy Confirmed Types. *Scientific Data*, 10:123, 2023.

[27] Yunliang Cai, Mark Landis, David T. Laidley, Anat Kornecki, Andrea Lum, and Shuo Li. Multi-Modal Vertebrae Recognition Using Transformed Deep Convolution Network. *Computerized Medical Imaging and Graphics*, 51:11–19, 2016.

[28] Yunliang Cai, Said Osman, Manas Sharma, Mark Landis, and Shuo Li. Multi-Modality Vertebra Recognition in Arbitrary Views Using 3D Deformable Hierarchical Model. *IEEE Transactions on Medical Imaging*, 34(8):1676–1693, 2015.

[29] Zhenjie Cao, Zhicheng Yang, Yuxing Tang, Yanbo Zhang, Mei Han, Jing Xiao, Jie Ma, and Peng Chang. Supervised Contrastive Pre-training for Mammographic Triage Screening Models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12907 of *Lecture Notes in Computer Science*, pages 129–139. Springer Nature Switzerland, 2021.

[30] Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. *IEEE Transactions on Medical Imaging*, 36(11):2355–2365, 2017.

[31] Eduardo Castro, Jose Costa Pereira, and Jaime S. Cardoso. Symmetry-Based Regularization in Deep Breast Cancer Screening. *Medical Image Analysis*, 83:102690, 2023.

[32] Heyou Chang, Shen Zhao, Hao Zheng, Yang Chen, and Shuo Li. Multi-Vertebrae Segmentation From Arbitrary Spine MR Images Under Global View. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12266 of *Lecture Notes in Computer Science*, pages 702–711. Springer, 2020.

[33] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled Representation Learning in Cardiac Image Analysis. *Medical Image Analysis*, 58:101535, 2019.

[34] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangle, Align and Fuse for Multimodal and Semi-Supervised Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(3):781–792, 2021.

[35] Cheng Chen, D. Belavy, and Guoyan Zheng. 3D Intervertebral Disc Localization and Segmentation From MR Images by Data-Driven Regression and Classification. In *Machine Learning in Medical Imaging (MLMI 2014)*, volume 8679 of *Lecture Notes in Computer Science*, pages 50–58. Springer International Publishing, 2014.

[36] Cheng Chen, Daniel Belavy, Weimin Yu, Chengwen Chu, Gabriele Armbrecht, Martin Bansmann, Dieter Felsenberg, and Guoyan Zheng. Localization and Segmentation of 3D Intervertebral Discs in MR Images by Data Driven Estimation. *IEEE Transactions on Medical Imaging*, 34(8):1719–1729, 2015.

[37] Hao Chen, Qi Dou, Xi Wang, Jing Qin, Jack C. Y. Cheng, and Pheng-Ann Heng. 3D Fully Convolutional Networks for Intervertebral Disc Localization and Segmentation. In *Medical Imaging and Augmented Reality (MIAR 2016)*, volume 9805 of *Lecture Notes in Computer Science*, pages 375–382. Springer, 2016.

[38] Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack C. Y. Cheng, and Pheng-Ann Heng. Automatic Localization and Identification of Vertebrae in Spine CT via a Joint Learning Model with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9349 of *Lecture Notes in Computer Science*, pages 515–522. Springer, 2015.

[39] Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao. Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model. *IEEE Transactions on Medical Imaging*, 39(2):387–399, 2020.

[40] Zixuan Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. CuNeRF: Cube-Based Neural Radiance Field for Zero-Shot Medical Image Arbitrary-Scale Super Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21185–21195, 2023.

[41] François Chollet et al. Keras. https://keras.io, 2015. Accessed: May 31 2023.

[42] Chengwen Chu, Daniel L. Belavý, Gabriele Armbrecht, Martin Bansmann, Dieter Felsenberg, and Guoyan Zheng. Fully Automatic Localization and Segmentation of 3D Vertebral Bodies from CT/MR Images via a Learning-Based Method. *PLOS ONE*, 10(11):e0143327, 2015.

[43] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation From Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9901 of *Lecture Notes in Computer Science*, pages 424–432. Springer International Publishing, 2016.

[44] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[45] Emily F. Conant, Brian L. Sprague, and Despina Kontos. Beyond BI-RADS Density: A Call for Quantification in the Breast Imaging Clinic. *Radiology*, 286(2):401–404, 2018.

[46] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active Shape Models – Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[47] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond Taylor, Sarath Bethapudi, Hubert P. H. Shum, and Chris G. Willcocks. MedNeRF: Medical Neural Radiance Fields for Reconstructing 3D-Aware CT-Projections From a Single X-Ray. In *Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3843–3848, 2022.

[48] Jason J. Corso, Raja' S. Alomari, and Vipin Chaudhary. Lumbar Disc Localization and Labeling With a Probabilistic Model on Both Pixel and Object Features. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2008*, volume 5241 of *Lecture Notes in Computer Science*, pages 202–210. Springer, 2008.

[49] Stefan Daenzer, Stefan Freitag, Sandra von Sachsen, Hanno Steinke, Mathias Groll, Jürgen Meixensberger, and Mario Leimert. VolHOG: A Volumetric Object Recognition Approach Based on Bivariate Histograms of Oriented Gradients for Vertebra Detection in Cervical Spine MRI. *Medical Physics*, 41(8):082305, 2014.

[50] Pabitra Das, Chandrajit Pal, Amit Acharyya, Amlan Chakrabarti, and Saumyajit Basu. Deep Neural Network for Automated Simultaneous Intervertebral Disc (IVDs) Identification and Segmentation of Multi-Modal MR Images. *Computer Methods and Programs in Biomedicine*, 205:106074, 2021.

[51] Stamatia V. Destounis, Amanda Santacroce, and Andrea Arieno. Update on Breast Density, Risk Estimation, and Supplemental Screening. *American Journal of Roentgenology*, 214(2):296–305, 2020.

[52] Yair Dgani, Hayit Greenspan, and Jacob Goldberger. Training a Neural Network Based on Unreliable Human Annotation of Medical Images. In *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pages 39–42. IEEE, 2018.

[53] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley. A Deep Learning Approach for the Analysis of Masses in Mammograms With Minimal User Intervention. *Medical Image Analysis*, 37:114–128, 2017.

[54] Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, Jack Kelly, Jeffrey de Fauw, Michael Heilman, Diogo Moitinho de Almeida, Brian McFee, Hendrik Weideman, Gábor Takács, Peter de Rivaz, Jon Crall, Gregory Sanders, Kashif Rasul, Cong Liu, Geoffrey French, and Jonas Degrave. Lasagne: First Release (v0.1). https://doi.org/10.5281/zenodo.27878, 2015. Accessed: May 31 2023.

[55] Kunio Doi. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[56] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. IVD-Net: Intervertebral Disc Localization and Segmentation in MRI With a Multi-Modal UNet. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2018)*, volume 11397 of *Lecture Notes in Computer Science*, pages 130–143. Springer International Publishing, 2019.

[57] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2019.

[58] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. PnP-AdaNet: Plug-and-Play Adversarial Domain Adaptation Network at Unpaired Cross-Modality Cardiac Segmentation. *IEEE Access*, 7:99065–99076, 2019.

[59] Issam El-Naqa, Yongyi Yang, Miles N. Wernick, Nikolas P. Galatsanos, and Robert M. Nishikawa. A Support Vector Machine Approach for Detection of Microcalcifications. *IEEE Transactions on Medical Imaging*, 21(12):1552–1563, 2002.

[60] Daniel Forsberg, Erik Sjöblom, and Jeffrey L. Sunshine. Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data. *Journal of Digital Imaging*, 30:406–412, 2017.

[61] Helen M. L. Frazer, Jennifer S. N. Tang, Michael S. Elliott, Katrina M. Kunicki, Brendan Hill, Ravishankar Karthik, Chun Fung Kwok, Carlos A. Peña-Solorzano, Yuanhong Chen, Chong Wang, Osamah Al-Qershi, Samantha K. Fox, Shuai Li, Enes Makalic, Tuong L. Nguyen, Daniel F. Schmidt, Prabhathi Basnayake Ralalage, Jocelyn F. Lippey, Peter Brotchie, John L. Hopper, Gustavo Carneiro, and Davis J. McCarthy. ADMANI: Annotated Digital Mammograms and Associated Non-Image Datasets. *Radiology: Artificial Intelligence*, 5(2):e220072, 2023.

[62] Frank Gaillard and Joshua Yap. Dixon Method: Reference Article, Radiopaedia.org. https://radiopaedia.org/articles/43255, Last update: 9 March 2023. Accessed: 31 May 2023.

[63] Ying'e Gao, Jingjing Lin, Yuzhuo Zhou, and Rongjin Lin. The Application of Traditional Machine Learning and Deep Learning Techniques in Mammography: A Review. *Frontiers in Oncology*, 13:1213045, 2023.

[64] Aimilia Gastounioti, Shyam Desai, Vinayak S. Ahluwalia, Emily F. Conant, and Despina Kontos. Artificial Intelligence in Mammographic Phenotyping of Breast Cancer Risk: A Narrative Review. *Breast Cancer Research*, 24:14, 2022.

[65] Krzysztof J. Geras, Ritse M. Mann, and Linda Moy. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. *Radiology*, 293(2):246–259, 2019.

[66] Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S. Gene Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks, 2018. arXiv preprint arXiv:1703.07047.

[67] Issachar Gilad and Moshe Nissan. Sagittal Evaluation of Elemental Geometrical Dimensions of Human Vertebrae. *Journal of Anatomy*, 143:115–120, 1985.

[68] Christina Gillmann, Noeska N. Smit, Eduard Gröller, Bernhard Preim, Anna Vilanova, and Thomas Wischgoll. Ten Open Challenges in Medical Visualization. *IEEE Computer Graphics and Applications*, 41(5):7–15, 2021.

[69] Ben Glocker, Johannes Feulner, Antonio Criminisi, David R. Haynor, and Ender Konukoglu. Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7512 of *Lecture Notes in Computer Science*, pages 590–598. Springer, 2012.

[70] Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi. Vertebrae Localization in Pathological Spine CT via Dense Classification From Sparse Annotations. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 262–270. Springer, 2013.

[71] Natalia Gorelik and Soterios Gyftopoulos. Applications of Artificial Intelligence in Musculoskeletal Imaging: From the Request to the Report. *Canadian Association of Radiologists' Journal*, 72(1):45–59, 2021.

[72] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022.

[73] Omid Haji Maghsoudi, Aimilia Gastounioti, Christopher Scott, Lauren Pantalone, Fang-Fang Wu, Eric A. Cohen, Stacey Winham, Emily F. Conant, Celine Vachon, and Despina Kontos. Deep-LIBRA: An Artificial-Intelligence Method for Robust Quantification of Breast Density with Independent Validation in Breast Cancer Risk Assessment. *Medical Image Analysis*, 73:102138, 2021.

[74] Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G. Wallis, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, and Kenneth C. Young. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. *Radiology: Artificial Intelligence*, 3(1):e200103, 2021.

[75] Zhongyi Han, Benzheng Wei, Ashley Mercado, Stephanie Leung, and Shuo Li. Spine-GAN: Semantic Segmentation of Multiple Spinal Structures. *Medical Image Analysis*, 50:23–35, 2018.

[76] Mohammad Hashir, Hadrien Bertrand, and Joseph Paul Cohen. Quantifying the Value of Lateral Views in Deep Learning for Chest X-Rays. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 288–303. PMLR, 2020.

[77] Michael D. Heath, Kevin W. Bowyer, Daniel B. Kopans, W. Philip Kegelmeyer Jr., Richard H. Moore, Kyong I. Chang, and S. Munishkumaran. Current Status of the Digital Database for Screening Mammography. In Nico Karssemeijer, Martin

Thijssen, Jan Hendriks, and Leon van Erning, editors, *Digital Mammography*, volume 13 of *Computational Imaging and Vision*, pages 457–460. Springer, 1998.

[78] Michael D. Heath, Kevin W. Bowyer, Daniel B. Kopans, Richard H. Moore, and W. Philip Kegelmeyer Jr. The Digital Database for Screening Mammography. In *Proceedings of the 5th International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

[79] Mattias P. Heinrich and Ozan Oktay. Accurate Intervertebral Disc Localisation and Segmentation in MRI Using Vantage Point Hough Forests and Multi-Atlas Fusion. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2016)*, volume 10182 of *Lecture Notes in Computer Science*, pages 77–84. Springer International Publishing, 2016.

[80] Haithem Hermessi, Olfa Mourali, and Ezzeddine Zagrouba. Multimodal Medical Image Fusion Review: Theoretical Background and Recent Advances. *Signal Processing*, 183:108036, 2021.

[81] Johannes Hofmanninger, Perkonigg, James A. Brink, Oleg Pianykh, Christian Herold, and Georg Langs. Dynamic Memory to Alleviate Catastrophic Forgetting in Continuous Learning Settings. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12262 of *Lecture Notes in Computer Science*, pages 359–368. Springer International Publishing, 2020.

[82] Johannes Hofmanninger, Florian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic Lung Segmentation in Routine Imaging is Primarily a Data Diversity Problem, Not a Methodology Problem. *European Radiology Experimental*, 4:50, 2020.

[83] Seyed-Parsa Hojjat, Ismail Ayed, Gregory J. Garvin, and Kumaradevan Punithakumar. Spine Labeling in MRI via Regularized Distribution Matching. *International Journal of Computer Assisted Radiology and Surgery*, 12(11):1911–1922, 2017.

[84] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J.W.L. Aerts. Artificial Intelligence in Radiology. *Nature Reviews. Cancer*, 18(8):500–510, 2018.

[85] Rui Hou, Yifan Peng, Lars J. Grimm, Yinhao Ren, Maciej A. Mazurowski, Jeffrey R. Marks, Lorraine M. King, Carlo C. Maley, E. Shelley Hwang, and Joseph Y. Lo. Anomaly Detection of Calcifications in Mammography Based on 11,000 Negative Cases. *IEEE Transactions on Biomedical Engineering*, 69(5):1639–1650, 2022.

[86] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017. arXiv preprint arxiv:1704.04861.

[87] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[88] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7319, 2017.

[89] Y. Huang, A. Uneri, C. K. Jones, X. Zhang, M. D. Ketcha, N. Aygun, P. A. Helm, and J. H. Siewerdsen. 3D Vertebrae Labeling in Spine CT: An Accurate, Memory-Efficient (Ortho2D) Framework. *Physics in Medicine and Biology*, 66(12):125020, 2021.

[90] Florian A. Huber and Roman Guggenberger. AI MSK Clinical Applications: Spine Imaging. *Skeletal radiology*, 51(2):279–291, 2022.

[91] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[92] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nature Methods*, 18(2):203–211, 2021.

[93] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. No New-Net. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2018)*, volume 11384 of *Lecture Notes in Computer Science*, pages 234–244. Springer Nature Switzerland, 2019.

[94] Pavel Izmailov, Dimitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 876–885, 2018.

[95] Roman Jakubicek, Jiri Chmelik, Jiri Jan, Petr Ourednicek, Lukas Lambert, and Giampaolo Gavelli. Learning-Based Vertebra Localization and Labeling in 3D CT Data of Possibly Incomplete and Pathological Spines. *Computer Methods and Programs in Biomedicine*, 183:105081, 2020.

[96] Amir Jamaludin, Meelis Lootus, Timor Kadir, and Andrew Zisserman. Automatic Intervertebral Discs Localization and Segmentation: A Vertebral Approach. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2015)*, volume 9402 of *Lecture Notes in Computer Science*, pages 97–103. Springer International Publishing, 2015.

[97] Alex Pappachen James and Belur V. Dasarathy. Medical Image Fusion: A Survey of the State of the Art. *Information Fusion*, 19:4–19, 2014.

[98] Xing Ji, Guoyan Zheng, Li Liu, and Dong Ni. Fully Automatic Localization and Segmentation of Intervertebral Disc from 3D Multi-Modality MR Images by Regression Forest and CNN. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2016)*, volume 10182 of *Lecture Notes in Computer Science*, pages 92–101. Springer International Publishing, 2016.

[99] Ana Jimenez-Pastor, Angel Alberich-Bayarri, Belen Fos-Guarinos, Fabio Garcia-Castro, David Garcia-Juan, Ben Glocker, and Luis Marti-Bonmati. Automated Vertebrae Localization and Identification by Decision Forests and Image-Based Refinement on Real-World CT Data. *La Radiologia Medica*, 125(1):48–56, 2020.

[100] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[101] Amira Jouirou, Abir Baâzaoui, and Walid Barhoumi. Multi-View Information Fusion in Mammograms: A Comprehensive Overview. *Information Fusion*, 52:308–321, 2019.

[102] Nicole Kaiser, Andreas Fieselmann, Sulaiman Vesal, Nishant Ravikumar, Ludwig Ritschl, Steffen Kappler, and Andreas K. Maier. Mammographic Breast Density Classification Using a Deep Neural Network: Assessment Based on Inter-Observer Variability. In *Proceedings of SPIE Medical Imaging*, volume 10952, pages 156–161. SPIE, 2019.

[103] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, 2016.

[104] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised Domain Adaptation in Brain Lesion Segmentation With Adversarial Networks. In *Information Processing in Medical Imaging*, volume 10265 of *Lecture Notes in Computer Science*, pages 597–609. Springer International Publishing, 2017.

[105] Neena Kapoor, Ronilda Lacson, and Ramin Khorasani. Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools. *Journal of the American College of Radiology*, 17(11):1363–1370, 2020.

[106] Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A Lifelong Learning Approach to Brain MR Segmentation Across Scanners and

Protocols. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11070 of *Lecture Notes in Computer Science*, pages 476–484. Springer, 2018.

[107] Nikola K. Kasabov. NeuCube: A Spiking Neural Network Architecture for Mapping, Learning and Understanding of Spatio-Temporal Brain Data. *Neural Networks*, 52:62–76, 2014.

[108] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation. *Medical Image Analysis*, 69:101950, 2021.

[109] Johannes Kehrer and Helwig Hauser. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, 2013.

[110] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. Changes in Cancer Detection and False-Positive Recall in Mammography Using Artificial Intelligence: A Retrospective, Multireader Study. *The Lancet. Digital Health*, 2(3):138–148, 2020.

[111] Thijs Kooi and Nico Karssemeijer. Classifying Symmetrical Differences and Temporal Change for the Detection of Malignant Masses in Mammography Using Deep Neural Networks. *Journal of Medical Imaging*, 4(4):044501, 2017.

[112] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions. *Medical Image Analysis*, 35:303–312, 2017.

[113] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009. Technical Report, University of Toronto.

[114] Xihe Kuang, Jason Pui Yin Cheung, Kwan-Yee K. Wong, Wai Yi Lam, Chak Hei Lam, Richard W. Choy, Christopher P. Cheng, Honghan Wu, Cao Yang, Kun Wang, Yang Li, and Teng Zhang. Spine-GFlow: A Hybrid Learning Framework for Robust Multi-Tissue Segmentation in Lumbar MRI Without Manual Annotation. *Computerized Medical Imaging and Graphics*, 99:102091, 2022.

[115] Trent Kyono, Fiona J. Gilbert, and Mihaela van der Schaar. Multi-View Multi-Task Learning for Improving Autonomous Mammogram Diagnosis. In *Proceedings of*

the 4th Machine Learning for Healthcare Conference, volume 106 of *Proceedings of Machine Learning Research*, pages 571–591. PMLR, 2019.

[116] Trent Kyono, Fiona J. Gilbert, and Mihaela van der Schaar. Improving Workflow Efficiency for Mammography Using Machine Learning. *Journal of the American College of Radiology*, 17:56–63, 2020.

[117] Dana Lahat, Tulay Adali, and Christian Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[118] Leslie R. Lamb, Constance D. Lehman, Aimilia Gastounioti, Emily F. Conant, and Manisha Bahl. Artificial Intelligence (AI) for Screening Mammography, From the AJR Special Series on AI Applications. *American Journal of Roentgenology*, 219(3):369–380, 2022.

[119] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[120] Juhun Lee and Robert M. Nishikawa. Identifying Women With Mammographically-Occult Breast Cancer Leveraging GAN-Simulated Mammograms. *IEEE Transactions on Medical Imaging*, 41(1):225–236, 2022.

[121] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research. *Scientific Data*, 4:170177, 2017.

[122] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) (Version 1) [Dataset], 2016. The Cancer Imaging Archive.

[123] Constance D. Lehman, Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, 290(1):52–58, 2019.

[124] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[125] Dimitrios Lenis, David Major, Maria Wimmer, Astrid Berg, Gert Sluiter, and Katja Bühler. Domain Aware Medical Image Classifier Interpretation by Counterfactual Impact Analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12261 of *Lecture Notes in Computer Science*, pages 315–325. Springer International Publishing, 2020.

[126] Nikolas Lessmann, Bram van Ginneken, Pim A. de Jong, and Ivana Išgum. Iterative Fully Convolutional Neural Networks for Automatic Vertebra Segmentation and Identification. *Medical Image Analysis*, 53:142–155, 2019.

[127] Chuanpu Li, Tianbao Liu, Zeli Chen, Shumao Pang, Liming Zhong, Qianjin Feng, and Wei Yang. SPA-ResUNet: Strip Pooling Attention ResUNet for Multi-Class Segmentation of Vertebrae and Intervertebral Discs. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[128] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C. Kot. Domain Generalization for Medical Imaging Classification With Linear-Dependency Regularization. In *Advances in Neural Information Processing Systems*, volume 33, pages 3118–3129. Curran Associates, Inc., 2020.

[129] Heyi Li, Dongdong Chen, William H. Nailon, Mike E. Davies, and David I. Laurenson. Dual Convolutional Neural Networks for Breast Mass Segmentation and Diagnosis in Mammography. *IEEE Transactions on Medical Imaging*, 41(1):3–13, 2022.

[130] Kang Li, Lequan Yu, and Pheng-Ann Heng. Domain-Incremental Cardiac Image Segmentation With Style-Oriented Replay and Domain-Sensitive Feature Whitening. *IEEE Transactions on Medical Imaging*, 42(3):570–581, 2023.

[131] Xiaomeng Li, Qi Dou, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Multi-Scale and Modality Dropout Learning for Intervertebral Disc Localization and Segmentation. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2016)*, volume 10182 of *Lecture Notes in Computer Science*, pages 85–91. Springer International Publishing, 2016.

[132] Xiaomeng Li, Qi Dou, Hao Chen, Chi-Wing Fu, Xiaojuan Qi, Daniel L. Belavý, Gabriele Armbrecht, Dieter Felsenberg, Guoyan Zheng, and Pheng-Ann Heng. 3D Multi-Scale FCN with Random Modality Voxel Dropout Learning for Intervertebral Disc Localization and Segmentation From Multi-modality MR Images. *Medical Image Analysis*, 45:41–54, 2018.

[133] Zheren Li, Zhiming Cui, Sheng Wang, Yuji Qi, Xi Ouyang, Qitian Chen, Yuezhi Yang, Zhong Xue, Dinggang Shen, and Jie-Zhi Cheng. Domain Generalization for Mammography Detection via Multi-Style and Multi-View Contrastive Learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12907 of *Lecture Notes in Computer Science*, pages 98–108. Springer Nature Switzerland, 2021.

[134] Haofu Liao, Addisu Mesfin, and Jiebo Luo. Joint Vertebrae Identification and Localization in Spinal CT Images by Combining Short- and Long-Range Contextual Information. *IEEE Transactions on Medical Imaging*, 37(5):1266–1275, 2018.

[135] Hans Liebl, David Schinz, Anjany Sekuboyina, Luca Malagutti, Maximilian T. Löffler, Amirhossein Bayat, Malek El Husseini, Giles Tetteh, Katharina Grau, Eva Niederreiter, Thomas Baum, Benedikt Wiestler, Bjoern Menze, Rickmer Braren, Claus Zimmer, and Jan S. Kirschke. A Computed Tomography Vertebral Segmentation Dataset With Anatomical Variations and Multi-Vendor Scanner Data. *Scientific Data*, 8(1):284, 2021.

[136] Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud A.A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

[137] Kangning Liu, Yiqiu Shen, Nan Wu, Jakub Chłędowski, Carlos Fernandez-Granda, and Krzysztof J. Geras. Weakly-Supervised High-Resolution Segmentation of Mammography Images for Breast Cancer Diagnosis. In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 451–472. PMLR, 2021.

[138] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning Disentangled Representations in the Imaging Domain. *Medical Image Analysis*, 80:102516, 2022.

[139] Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu. Act Like a Radiologist: Towards Reliable Multi-View Correspondence Reasoning for Mammogram Mass Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5947–5961, 2022.

[140] Joe Logan, Paul J. Kennedy, and Daniel Catchpoole. A Review of the Machine Learning Datasets in Mammography, their Adherence to the FAIR Principles and the Outlook for the Future. *Scientific Data*, 10:595, 2023.

[141] Meelis Lootus, Timor Kadir, and Andrew Zisserman. Vertebrae Detection and Labelling in Lumbar MR Images. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2014)*, volume 17 of *Lecture Notes in Computational Vision and Biomechanics*, pages 219–230. Springer International Publishing, 2014.

[142] Meelis Lootus, Timor Kadir, and Andrew Zisserman. Automated Radiological Grading of Spinal MRI. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging (CSI 2015)*, volume 20 of *Lecture Notes in Computational Vision and Biomechanics*, pages 119–130. Springer International Publishing, 2015.

[143] William Lotter, Abdul Rahman Diab, Bryan Haslam, Jiye G. Kim, Giorgia Grisot, Eric Wu, Kevin Wu, Jorge Onieva Onieva, Yun Boyer, Jerrold L. Boxerman, Meiyun Wang, Mack Bandler, Gopal R. Vijayaraghavan, and A. Gregory Sorensen. Robust Breast Cancer Detection in Mammography and Digital Breast Tomosynthesis Using

an Annotation-Efficient Deep Learning Approach. *Nature Medicine*, 27(2):244–249, 2021.

[144] William Lotter, Greg Sorensen, and David Cox. A Multi-Scale CNN and Curriculum Learning Strategy for Mammogram Classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS 2017)*, volume 10553 of *Lecture Notes in Computer Science*, pages 169–177. Springer International Publishing, 2017.

[145] Jen-Tang Lu, Stefano Pedemonte, Bernardo Bizzo, Sean Doyle, Katherine P. Andriole, Mark H. Michalski, R. Gilberto Gonzalez, and Stuart R. Pomerantz. Deep Spine: Automated Lumbar Vertebral Segmentation, Disc-Level Designation, and Spinal Stenosis Grading Using Deep Learning. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 403–419. PMLR, 2018.

[146] Luyang Luo, Xi Wang, Yi Lin, Xiaoqi Ma, Andong Tan, Ronald Chan, Varut Vardhanabhuti, Winnie CW Chu, Kwang-Ting Cheng, and Hao Chen. Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions. *IEEE Reviews in Biomedical Engineering*, 2024. to appear.

[147] Jiechao Ma, Xiang Li, Hongwei Li, Ruixuan Wang, Bjoern Menze, and Wei-Shi Zheng. Cross-View Relation Networks for Mammogram Mass Detection. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pages 8632–8638, 2021.

[148] Frederick Maes, David Robben, Dirk Vandermeulen, and Paul Suetens. The Role of Medical Image Computing and Machine Learning in Healthcare. In Paul R. Algra, Sergey Morozov, and Erik R. Ranschaert, editors, *Artificial Intelligence in Medical Imaging*, pages 9–23. Springer International Publishing, 2019.

[149] David Major, Jiří Hladůvka, Florian Schulze, and Katja Bühler. Automated Landmarking and Labeling of Fully and Partially Scanned Spinal Columns in CT Images. *Medical Image Analysis*, 17(8):1151–1163, 2013.

[150] David Major, Dimitrios Lenis, Maria Wimmer, Astrid Berg, Theresa Neubauer, and Katja Bühler. On the Importance of Domain Awareness in Classifier Interpretations in Medical Imaging. *IEEE Transactions on Medical Imaging*, 42(8):2286–2298, 2023.

[151] David Major, Dimitrios Lenis, Maria Wimmer, Gert Sluiter, Astrid Berg, and Katja Bühler. Interpreting Medical Image Classifiers by Optimization Based Counterfactual Impact Analysis. In *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1096–1100. IEEE, 2020.

[152] Ritse M. Mann, Regina Hooley, Richard G. Barr, and Linda Moy. Novel Approaches to Screening for Breast Cancer. *Radiology*, 297(2):266–285, 2020.

[153] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Gordon H. Bower, editor, *Advances in Research and Theory*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, Inc., 1989.

[154] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey de Fauw, and Shravya Shetty. International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788):89–94, 2020.

[155] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019.

[156] B. Michael Kelm, Michael Wels, S. Kevin Zhou, Sascha Seifert, Michael Suehling, Yefeng Zheng, and Dorin Comaniciu. Spine Detection in CT and MR Using Iterated Marginal Space Learning. *Medical Image Analysis*, 17(8):1283–1292, 2013.

[157] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer International Publishing, 2020.

[158] Aly A. Mohamed, Wendie A. Berg, Hong Peng, Yahong Luo, Rachel C. Jankowitz, and Shandong Wu. A Deep Learning Method for Classifying Mammographic Breast Density Categories. *Medical Physics*, 45(1):314–321, 2018.

[159] Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit Neural Representation in Medical Imaging: A Comparative Survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2381–2391, 2023.

[160] Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida, and Nico Karssemeijer. Automatic Microcalcification Detection in Multi-Vendor Mammography Using Convolutional Neural Networks. In *Breast Imaging*, volume 9699 of *Lecture Notes in Computer Science*, pages 35–42. Springer International Publishing, 2016.

[161] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. INbreast: Toward a Full-Field Digital Mammographic Database. *Academic Radiology*, 19(2):236–248, 2012.

[162] Theresa Neubauer, Maria Wimmer, Astrid Berg, David Major, Dimitrios Lenis, Thomas Beyer, Jelena Saponjski, and Katja Bühler. Soft Tissue Sarcoma Co-Segmentation in Combined MRI and PET/CT Data. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (ML-CDS 2020/CLIP 2020)*, volume 12445 of *Lecture Notes in Computer Science*, pages 97–105. Springer Nature Switzerland, 2020.

[163] Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, and Van Vu. VinDr-Mammo: A Large-Scale Benchmark Dataset for Computer-Aided Diagnosis in Full-Field Digital Mammography. *Scientific Data*, 10:277, 2023.

[164] Alexey A. Novikov, Dimitrios Lenis, David Major, Jiří Hladůvka, Maria Wimmer, and Katja Bühler. Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs. *IEEE Transactions on Medical Imaging*, 37(8):1865–1876, 2018.

[165] Alexey A. Novikov, David Major, Maria Wimmer, Dimitrios Lenis, and Katja Bühler. Deep Sequential Segmentation of Organs in Volumetric Medical Scans. *IEEE Transactions on Medical Imaging*, 38(5):1207–1215, 2019.

[166] OECD. Health Care Utilisation: Diagnostic Exams. https://stats.oecd.org/index.aspx?queryid=30160, Last update: 4 July 2022. Accessed: 11 July 2022.

[167] Ayse Betul Oktay and Yusuf Sinan Akgul. Simultaneous Localization of Lumbar Vertebrae and Intervertebral Discs with SVM-Based MRF. *IEEE Transactions on Biomedical Engineering*, 60(9):2375–2383, 2013.

[168] Arnau Oliver, Jordi Freixenet, Robert Martí, Josep Pont, Elsa Pérez, Erika R. E. Denton, and Reyer Zwiggelaar. A Novel Breast Tissue Density Classification Methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65, 2008.

[169] William C. Ou, Dogan Polat, and Basak E. Dogan. Deep Learning in Breast Radiology: Current Progress and Future Directions. *European Radiology*, 31(7):4872–4885, 2021.

[170] Shumao Pang, Chunlan Pang, Zhihai Su, Liyan Lin, Lei Zhao, Yangfan Chen, Yujia Zhou, Hai Lu, and Qianjin Feng. DGMSNet: Spine Segmentation for MR Image by a Detection-Guided Mixed-Supervised Segmentation Network. *Medical Image Analysis*, 75:102261, 2022.

[171] Manohar M. Panjabi, Vijay Goel, Thomas Oxland, Koichiro Takata, Joanne Duranceau, Martin Krag, and Mark Price. Human Lumbar Vertebrae: Quantitative Three-Dimensional Anatomy. *Spine*, 17(3):299–306, 1992.

[172] Manohar M. Panjabi, Koichiro Takata, Vijay Goel, Dale Federico, Thomas Oxland, Joanne Duranceau, and Martin Krag. Thoracic Human Vertebrae: Quantitative Three-Dimensional Anatomy. *Spine*, 16(8):888–901, 1991.

[173] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[174] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Coarse to Fine Vertebrae Localization and Segmentation With SpatialConfiguration-Net and U-Net. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pages 124–133. SCITEPRESS - Science and Technology Publications, 2020.

[175] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[176] Matthias Perkonigg, Johannes Hofmanninger, Christian J. Herold, James A. Brink, Oleg Pianykh, Helmut Prosch, and Georg Langs. Dynamic Memory to Alleviate Catastrophic Forgetting in Continual Learning With Medical Imaging. *Nature Communications*, 12:5678, 2021.

[177] John H. Phan, Chang F. Quo, Chihwen Cheng, and May Dongmei Wang. Multiscale Integration of -Omic, Imaging, and Clinical Data in Biomedical Informatics. *IEEE Reviews in Biomedical Engineering*, 5:74–87, 2012.

[178] Oleg S. Pianykh, Georg Langs, Marc Dewey, Dieter R. Enzmann, Christian J. Herold, Stefan O. Schoenberg, and James A. Brink. Continuous Learning AI in Radiology: Implementation Principles and Early Applications. *Radiology*, 297:6–14, 2020.

[179] Bernhard Preim and Charl Botha. Acquisition of Medical Image Data. In *Visual Computing for Medicine*, pages 15–67. Elsevier, 2014.

[180] Bernhard Preim and Charl Botha. An Introduction to Medical Visualization in Clinical Practice. In *Visual Computing for Medicine*, pages 69–110. Elsevier, 2014.

[181] Biao Qu, Jianpeng Cao, Chen Qian, Jinyu Wu, Jianzhong Lin, Liansheng Wang, Lin Ou-Yang, Yongfa Chen, Liyue Yan, Qing Hong, Gaofeng Zheng, and Xiaobo Qu.

Current Development and Prospects of Deep Learning in Spine Image Analysis: A Literature Review. *Quantitative Imaging in Medicine and Surgery*, 12(6):3454–3479, 2022.

[182] Renata Georgia Raidou. Visual Analytics for the Representation, Exploration, and Analysis of High-Dimensional, Multi-Faceted Medical Data. In Paul M. Rea, editor, *Biomedical Visualisation*, volume 1138 of *Advances in Experimental Medicine and Biology*, pages 137–162. Springer, 2019.

[183] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in Health and Medicine. *Nature Medicine*, 28:31–38, 2022.

[184] Marko Rak and Klaus D. Tönnies. On Computerized Methods for Spine Analysis in MRI: A Systematic Review. *International Journal of Computer Assisted Radiology and Surgery*, 11(8):1445–1465, 2016.

[185] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015.

[186] Yinhao Ren, Jiafeng Lu, Zisheng Liang, Lars J. Grimm, Connie Kim, Michael Taylor-Cho, Sora Yoon, Jeffrey R. Marks, and Joseph Y. Lo. Retina-Match: Ipsilateral Mammography Lesion Matching in a Single Shot Detection Pipeline. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12905 of *Lecture Notes in Computer Science*, pages 345–354. Springer International Publishing, 2021.

[187] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[188] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and Classifying Lesions in Mammograms With Deep Learning. *Scientific Reports*, 8:4165, 2018.

[189] Michael L. Richardson, Elisabeth R. Garwood, Yueh Lee, Matthew D. Li, Hao S. Lo, Arun Nagaraju, Xuan V. Nguyen, Linda Probyn, Prabhakar Rajiah, Jessica Sin, Ashish P. Wasnik, and Kali Xu. Noninterpretive Uses of Artificial Intelligence in Radiology. *Academic Radiology*, 28(9):1225–1235, 2021.

[190] Nicolas Robitaille, Abderazzak Mouiha, Burt Crépeault, Fernando Valdivia, and Simon Duchesne. Tissue-Based MRI Intensity Standardization: Application to Multicentric Datasets. *International Journal of Biomedical Imaging*, 2012:347120, 2012.

[191] Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, and Ritse M. Mann. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology*, 290(2):305–314, 2019.

[192] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H. Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, Matthew G. Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M. Mann, and Ioannis Sechopoulos. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison with 101 Radiologists. *Journal of the National Cancer Institute*, 111(9):916–922, 2019.

[193] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Jonas Teuwen, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H. Helbich, Margarita Chevalier, Thomas Mertelmeier, Matthew G. Wallis, Ingvar Andersson, Sophia Zackrisson, Ioannis Sechopoulos, and Ritse M. Mann. Can We Reduce the Workload of Mammographic Screening by Automatic Identification of Normal Exams With Artificial Intelligence? A Feasibility Study. *European Radiology*, 29(9):4825–4832, 2019.

[194] Ravi K. Samala, Heang-Ping Chan, Lubomir M. Hadjiiski, Mark A. Helvie, Kenny H. Cha, and Caleb D. Richter. Multi-Task Transfer Learning Deep Convolutional Neural Network: Application to Computer-Aided Diagnosis of Breast Cancer on Mammograms. *Physics in Medicine and Biology*, 62(23):8894–8908, 2017.

[195] Nagwan Abdel Samee, Amel A. Alhussan, Vidan Fathi Ghoneim, Ghada Atteia, Reem Alkanhel, Mugahed A. Al-Antari, and Yasser M. Kadah. A Hybrid Deep Transfer Learning of CNN-Based LR-PCA for Breast Lesion Diagnosis via Medical Breast Mammograms. *Sensors*, 22(13):4938, 2022.

[196] Nagwan Abdel Samee, Ghada Atteia, Souham Meshoul, Mugahed A. Al-Antari, and Yasser M. Kadah. Deep Learning Cascaded Feature Selection Framework for Breast Cancer Classification: Hybrid CNN with Univariate-Based Approach. *Mathematics*, 10(19):3631, 2022.

[197] Jörg Sander, Bob D. de Vos, Steffen Bruns, Nils Planken, Max A. Viergever, Tim Leiner, and Ivana Išgum. Reconstruction and Completion of High-Resolution 3D Cardiac Shapes using Anisotropic CMRI Segmentations and Continuous Implicit Neural Representations. *Computers in Biology and Medicine*, 164:107266, 2023.

[198] Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezso Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, Jiashi Feng, Mengling Feng, Hyo-Eun Kim, Francisco Albiol, Alberto Albiol, Stephen Morrell, Zbigniew Wojna, Mehmet Eren Ahsen, Umar Asif, Antonio Jimeno Yepes, Shivanthan Yohanandan, Simona Rabinovici-Cohen, Darvin Yi, Bruce Hoff, Thomas Yu, Elias Chaibub Neto, Daniel L. Rubin, Peter Lindholm, Laurie R. Margolies,

Russell Bailey McBride, Joseph H. Rothstein, Weiva Sieh, Rami Ben-Ari, Stefan Harrer, Andrew Trister, Stephen Friend, Thea Norman, Berkman Sahiner, Fredrik Strand, Justin Guinney, Gustavo Stolovitzky, Lester Mackey, Joyce Cahoon, Li Shen, Jae Ho Sohn, Hari Trivedi, Yiqiu Shen, Ljubomir Buturovic, Jose Costa Pereira, Jaime S. Cardoso, Eduardo Castro, Karl Trygve Kalleberg, Obioma Pelka, Imane Nedjar, Krzysztof J. Geras, Felix Nensa, Ethan Goan, Sven Koitka, Luis Caballero, David D. Cox, Pavitra Krishnaswamy, Gaurav Pandey, Christoph M. Friedrich, Dimitri Perrin, Clinton Fookes, Bibo Shi, Gerard Cardoso Negrie, Michael Kawczynski, Kyunghyun Cho, Can Son Khoo, Joseph Y. Lo, A. Gregory Sorensen, and Hwejin Jung. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*, 3(3):e200265, 2020.

[199] Stefan Schmidt, Jörg Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian Dries, Daniel Bystrov, and Christoph Schnörr. Spine Detection and Labeling Using a Parts-Based Graphical Model. In *Information Processing in Medical Imaging*, volume 4584 of *Lecture Notes in Computer Science*, pages 122–133. Springer, 2007.

[200] Florian Schulze, David Major, and Katja Bühler. Fast and Memory Efficient Feature Detection Using Multiresolution Probabilistic Boosting Trees. *Journal of WSCG*, 19(1):33–40, 2011.

[201] Ioannis Sechopoulos, Jonas Teuwen, and Ritse Mann. Artificial Intelligence for Breast Cancer Detection in Mammography and Digital Breast Tomosynthesis: State of the Art. *Seminars in Cancer Biology*, 72:214–225, 2021.

[202] Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olaguer, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Kevin Brown, Alexandre Kirszenberg, Élodie Puybareau, Di Chen, Yiwei Bai, Brandon H. Rapazzo, Timyoas Yeah, Amber Zhang, Shangliang Xu, Feng Hou, Zhiqiang He, Chan Zeng, Zheng Xiangshang, Xu Liming, Tucker J. Netherton, Raymond P. Mumme, Laurence E. Court, Zixun Huang, Chenhang He, Li-Wen Wang, Sai Ho Ling, Lê Duy Huỳnh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. VerSe: A Vertebrae Labelling and Segmentation Benchmark for Multi-Detector CT Images. *Medical Image Analysis*, 73:102166, 2021.

[203] Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitsch, Jan S. Kirschke, and Bjoern H. Menze. Btrfly Net: Vertebrae

Labelling with Energy-Based Adversarial Learning of Local Spine Prior. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11073 of *Lecture Notes in Computer Science*, pages 649–657. Springer International Publishing, 2018.

[204] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sanchez, and Bram van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[205] Yaniv Shachor, Hayit Greenspan, and Jacob Goldberger. A Mixture of Views Network With Applications to Multi-View Medical Imaging. *Neurocomputing*, 374:1–9, 2020.

[206] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, 9(1):12495, 2019.

[207] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Globally-Aware Multiple Instance Classifier for Breast Cancer Screening. In *Machine Learning in Medical Imaging (MLMI 2019)*, volume 11861 of *Lecture Notes in Computer Science*, pages 18–26. Springer, 2019.

[208] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An Interpretable Classifier for High-Resolution Breast Cancer Screening Images Utilizing Weakly Supervised Localization. *Medical Image Analysis*, 68:10908, 2021.

[209] Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, and Dosik Hwang. COSMOS: Cross-Modality Unsupervised Domain Adaptation for 3D Medical Image Segmentation Based on Target-Aware Domain Translation and Iterative Self-Training, 2022. arXiv preprint arXiv:2203.16557.

[210] Russell T. Shinohara, Elizabeth M. Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J. Mateen, Peter A. Calabresi, Samson Jarso, Dzung L. Pham, Daniel S. Reich, and Ciprian M. Crainiceanu. Statistical Normalization Techniques for Magnetic Resonance Imaging. *NeuroImage: Clinical*, 6:9–19, 2014.

[211] Xin Shu, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep Neural Networks With Region-Based Pooling Structures for Mammographic Image Classification. *IEEE Transactions on Medical Imaging*, 39(6):2246–2255, 2020.

[212] Edward A. Sickles, Carl J. D'Orsi, Lawrence W. Bassett, et al. ACR BI-RADS® Mammography. In *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. American College of Radiology, 2013.

[213] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit Neural Representations With Periodic Activation Functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020.

[214] Shelly Soffer, Avi Ben-Cohen, Orit Shimon, Michal Marianne Amitai, Hayit Greenspan, and Eyal Klang. Convolutional Neural Networks for Radiologic Images: A Radiologist´s Guide. *Radiology*, 290(3):590–606, 2019.

[215] Spineweb Global. SpineWeb: Collaborative Platform for Research on Spine Imaging and Image Analysis. http://spineweb.digitalimaginggroup.ca, Last update: 13 May 2016. Accessed: 2 April 2023.

[216] Brian L. Sprague, Emily F. Conant, Tracy Onega, Michael P. Garcia, Elisabeth F. Beaber, Sally D. Herschorn, Constance D. Lehman, Anna N. A. Tosteson, Ronilda Lacson, Mitchell D. Schnall, Despina Kontos, Jennifer S. Haas, Donald L. Weaver, and William E. Barlow. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice: Findings From a Multicenter Observational Study. *Annals of Internal Medicine*, 165(7):457–464, 2016.

[217] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[218] Shikhar Srivastava, Mohammad Yaqub, Karthik Nandakumar, Zongyuan Ge, and Dwarikanath Mahapatra. Continual Domain Incremental Learning for Chest X-Ray Classification in Low-Resource Clinical Settings. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health (DART 2021/FAIR 2021)*, volume 12968 of *Lecture Notes in Computer Science*, pages 226–238. Springer International Publishing, 2021.

[219] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal Deep Learning for Biomedical Data Fusion: A Review. *Briefings in Bioinformatics*, 23(2):1–15, 2022.

[220] Darko Štern, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. Automated Detection of Spinal Centrelines, Vertebral Bodies and Intervertebral Discs in CT and MR Images of Lumbar Spine. *Physics in Medicine and Biology*, 55(1):247–264, 2009.

[221] Amin Suzani, Alexander Seitel, Yuan Liu, Sidney Fels, Robert N. Rohling, and Purang Abolmaesumi. Fast Automatic Vertebrae Detection and Localization in Pathological CT Scans - A Deep Learning Approach. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 678–686. Springer, 2015.

[222] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[223] Tao Tan, Alejandro Rodriguez-Ruiz, Tianyu Zhang, Lin Xu, Regina G. H. Beets-Tan, Yingzhao Shen, Nico Karssemeijer, Jun Xu, Ritse M. Mann, and Lingyun Bao. Multi-Modal Artificial Intelligence for the Combination of Automated 3D Breast Ultrasound and Mammograms in a Population of Women with Predominantly Dense Breasts. *Insights into Imaging*, 14:10, 2023.

[224] Rong Tao, Wenyong Liu, and Guoyan Zheng. Spine-Transformers: Vertebra Labeling and Segmentation in Arbitrary Field-Of-View Spine CTs via 3D Transformers. *Medical Image Analysis*, 75:102258, 2022.

[225] Mickael Tardy and Diana Mateus. Looking for Abnormalities in Mammograms With Self- and Weakly Supervised Reconstruction. *IEEE Transactions on Medical Imaging*, 40(10):2711–2722, 2021.

[226] Mickael Tardy and Diana Mateus. Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms. *Frontiers in Radiology*, 1:796078, 2022.

[227] Tom van Sonsbeek, Pardiss Danaei, Delaram Behnami, Mohammad Hossein Jafari, Parisa Asgharzadeh, Robert Rohling, and Purang Abolmaesumi. End-To-End Vertebra Localization and Level Detection in Weakly Labelled 3D Spinal MR Using Cascaded Neural Networks. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pages 1178–1182. IEEE, 2019.

[228] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human-Computer Collaboration for Skin Cancer Recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[229] Lazaros Tsochatzidis, Panagiota Koutla, Lena Costaridou, and Ioannis Pratikakis. Integrating Segmentation Information into CNN for Breast Cancer Diagnosis of Mammographic Masses. *Computer Methods and Programs in Biomedicine*, 200:105913, 2021.

[230] Gido M. van de Ven and Andreas S. Tolias. Three Scenarios for Continual Learning, 2019. arXiv preprint arXiv:1904.07734.

[231] Malinda Vania and Deukhee Lee. Intervertebral Disc Instance Segmentation Using a Multistage Optimization Mask-RCNN (MOM-RCNN). *Journal of Computational Design and Engineering*, 8(4):1023–1036, 2021.

[232] Srinivasan Vedantham, Mohammed Salman Shazeeb, Alan Chiang, and Gopal R. Vijayaraghavan. Artificial Intelligence in Breast X-Ray Imaging. *Seminars in Ultrasound, CT, and MRI*, 44(1):2–7, 2023.

[233] Paul Viola and Michael Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518. 2001.

[234] Ricky Walsh and Mickael Tardy. A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer. *Diagnostics*, 13(1):67, 2022.

[235] Alexander J. T. Wanders, Willem Mees, Petra A. M. Bun, Natasja Janssen, Alejandro Rodríguez-Ruiz, Mehmet Ufuk Dalmış, Nico Karssemeijer, Carla H. van Gils, Ioannis Sechopoulos, Ritse M. Mann, and Cornelis Jan van Rooden. Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms. *Radiology*, 303(2):269–275, 2022.

[236] Chong Wang, Yuanhong Chen, Yuyuan Liu, Yu Tian, Fengbei Liu, Davis J. McCarthy, Michael Elliott, Helen Frazer, and Gustavo Carneiro. Knowledge Distillation to Ensemble Global and Interpretable Prototype-Based Mammogram Classification Models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13433 of *Lecture Notes in Computer Science*, pages 14–24. Springer Nature Switzerland, 2022.

[237] Juan Wang and Yongyi Yang. A Context-Sensitive Deep Learning Approach for Microcalcification Detection in Mammograms. *Pattern Recognition*, 78:12–22, 2018.

[238] Junxia Wang, Yuanjie Zheng, Jun Ma, Xinmeng Li, Chongjing Wang, James Gee, Haipeng Wang, and Wenhui Huang. Information Bottleneck-Based Interpretable Multitask Network for Breast Cancer Classification and Segmentation. *Medical Image Analysis*, 83:102687, 2023.

[239] Yan Wang, Yangqin Feng, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep Adversarial Domain Adaptation for Breast Cancer Screening from Mammograms. *Medical Image Analysis*, 73:102147, 2021.

[240] Yan Wang, Zizhou Wang, Yangqin Feng, and Lei Zhang. WDCCNet: Weighted Double-Classifier Constraint Neural Network for Mammographic Image Classification. *IEEE Transactions on Medical Imaging*, 41(3):559–570, 2022.

[241] Zhijie Wang, Xiantong Zhen, KengYeow Tay, Said Osman, Walter Romano, and Shuo Li. Regression Segmentation for M3 Spinal Images. *IEEE Transactions on Medical Imaging*, 34(8):1640–1648, 2015.

[242] Olena Weaver and Jessica W. T. Leung. Biomarkers and Imaging of Breast Cancer. *American Journal of Roentgenology*, 210(2):271–278, 2018.

[243] Tao Wei, Angelica I. Aviles-Rivero, Shuo Wang, Yuan Huang, Fiona J. Gilbert, Carola-Bibiane Schönlieb, and Chang Wen Chen. Beyond Fine-Tuning: Classifying High Resolution Mammograms using Function-Preserving Transformations. *Medical Image Analysis*, 82:102618, 2022.

[244] Neil I. Weisenfeld and Simon K. Warfield. Normalization of Joint Image-Intensity Statistics in MRI Using the Kullback-Leibler Divergence. In *IEEE 2nd International Symposium on Biomedical Imaging (ISBI)*, pages 101–104. IEEE, 2004.

[245] Christopher P. Wild, Elisabete Weiderpass, and Bernard W. Stewart. *World Cancer Report 2020: Cancer Research for Cancer Prevention*. International Agency for Research on Cancer, Lyon, 2020.

[246] Maria Wimmer. Semi-Automatic Spine Labeling on T1- and T2-Weighted MRI Volume Data. Master's thesis, TU Wien, Vienna, 2015.

[247] Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. Local Entropy-Optimized Texture Models for Semi-Automatic Spine Labeling in Various MRI Protocols. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 155–159. IEEE, 2016.

[248] Maria Wimmer, David Major, Alexey A. Novikov, and Katja Bühler. Fully Automatic Cross-Modality Localization and Labeling of Vertebral Bodies and Intervertebral Discs in 3D Spinal Images. *International Journal of Computer Assisted Radiology and Surgery*, 13(10):1591–1603, 2018.

[249] Maria Wimmer, Gert Sluiter, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, and Katja Bühler. Multi-Task Fusion for Improving Mammography Screening Data Classification. *IEEE Transactions on Medical Imaging*, 41(4):937–950, 2022.

[250] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12266 of *Lecture Notes in Computer Science*, pages 712–722. Springer International Publishing, 2020.

[251] Nan Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim, Stacey Wolfson, Linda Moy, and Kyunghyun Cho. Breast Density Classification with Deep Convolutional Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6682–6686. IEEE, 2018.

[252] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng L.Y. Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon

Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020.

[253] Qing Wu, Yuwei Li, Yawen Sun, Yan Zhou, Hongjiang Wei, Jingyi Yu, and Yuyao Zhang. An Arbitrary Scale Super-Resolution Approach for 3D MR Images via Implicit Neural Representation. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1004–1015, 2023.

[254] Qing Wu, Yuwei Li, Lan Xu, Ruiming Feng, Hongjiang Wei, Qing Yang, Boliang Yu, Xiaozhao Liu, Jingyi Yu, and Yuyao Zhang. IREM: High-Resolution Magnetic Resonance Image Reconstruction via Implicit Neural Representation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12906 of *Lecture Notes in Computer Science*, pages 65–74. Springer Nature Switzerland, 2021.

[255] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 292(1):60–66, 2019.

[256] Adam Yala, Peter G. Mikhal, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Lian Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward Robust Mammography-Based Models for Breast Cancer Risk. *Science Translational Medicine*, 13(578):eaba4373, 2021.

[257] Zhicheng Yang, Zhenjie Cao, Yanbo Zhang, Yuxing Tang, Xiaohui Lin, Rushan Ouyang, Mingxiang Wu, Mei Han, Jing Xiao, Lingyun Huang, Shibin Wu, Peng Chang, and Jie Ma. MommiNet-v2: Mammographic Multi-View Mass Identification Networks. *Medical Image Analysis*, 73:102204, 2021.

[258] Jung Hyun Yoon and Eun-Kyung Kim. Deep Learning-Based Artificial Intelligence for Mammography. *Korean Journal of Radiology*, 22(8):1225–1239, 2021.

[259] Kihyun You, Suho Lee, Kyuhee Jo, Eunkyung Park, Thijs Kooi, and Hyeonseob Nam. Intra-Class Contrastive Learning Improves Computer Aided Diagnosis of Breast Cancer in Mammography. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13433 of *Lecture Notes in Computer Science*, pages 55–64. Springer Nature Switzerland, 2022.

[260] Sebastian Zambal, Katja Bühler, and Jirí Hladůvka. Entropy-Optimized Texture Models. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, volume 5242 of *Lecture Notes in Computer Science*, pages 213–221. Springer, 2008.

[261] Guodong Zeng, Daniel Belavy, Shuo Li, and Guoyan Zheng. Evaluation and Comparison of Automatic Intervertebral Disc Localization and Segmentation Methods

With 3D Multi-Modality MR Images: A Grand Challenge. In *Computational Methods and Clinical Applications for Spine Imaging (CSI 2018)*, volume 11397 of *Lecture Notes in Computer Science*, pages 163–171. Springer International Publishing, 2019.

[262] Yiqiang Zhan, Bing Jian, Dewan Maneesh, and Xiang Sean Zhou. Cross-Modality Vertebrae Localization and Labeling Using Learning-Based Approaches. In *Spinal Imaging and Image Analysis*, volume 18 of *Lecture Notes in Computational Vision and Biomechanics*, pages 301–322. Springer International Publishing, 2015.

[263] Dong Zhang, Bo Chen, and Shuo Li. Sequential Conditional Reinforcement Learning for Simultaneous Vertebral Body Detection and Segmentation With Modeling the Spine Anatomy. *Medical Image Analysis*, 67:101861, 2021.

[264] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540, 2020.

[265] Yu-Dong Zhang, Zhengchao Dong, Shui-Hua Wang, Xiang Yu, Xujing Yao, Qinghua Zhou, Hua Hu, Min Li, Carmen Jiménez-Mesa, Javier Ramirez, Francisco J. Martinez, and Juan Manuel Gorriz. Advances in Multimodal Data Fusion in Neuroimaging: Overview, Challenges, and Novel Orientation. *Information Fusion*, 64:149–187, 2020.

[266] Shen Zhao, Xi Wu, Bo Chen, and Shuo Li. Automatic Vertebrae Recognition From Arbitrary Spine MRI Images by a Category-Consistent Self-Calibration Detection Framework. *Medical Image Analysis*, 67:101826, 2021.

[267] Guoyan Zheng, Chengwen Chu, Daniel L. Belavý, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, Ben Glocker, Hao Chen, Qi Dou, Pheng-Ann Heng, Chunliang Wang, Daniel Forsberg, Aleš Neubert, Jürgen Fripp, Martin Urschler, Darko Stern, Maria Wimmer, Alexey A. Novikov, Hui Cheng, Gabriele Armbrecht, Dieter Felsenberg, and Shuo Li. Evaluation and Comparison of 3D Intervertebral Disc Localization and Segmentation Methods for 3D T2 MR Data: A Grand Challenge. *Medical Image Analysis*, 35:327–344, 2017.

[268] Yutong Zhong, Yan Piao, Baolin Tan, and Jingxin Liu. A Multi-Task Fusion Model based on a Residual-Multi-Layer Perceptron Network for Mammographic Breast Cancer Screening. *Computer Methods and Programs in Biomedicine*, 247:108101, 2024.

[269] Zisha Zhong, Yusung Kim, Kristin Plichta, Bryan G. Allen, Leixin Zhou, John Buatti, and Xiaodong Wu. Simultaneous Cosegmentation of Tumors in PET-CT Images Using Deep Fully Convolutional Networks. *Medical Physics*, 46(2):619–633, 2019.

[270] Tongxue Zhou, Su Ruan, and Stéphane Canu. A Review: Deep Learning for Medical Image Segmentation Using Multi-Modality Fusion. *Array*, 3-4:10004, 2019.

[271] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep Multi-Instance Networks With Sparse Label Assignment for Whole Mammogram Classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, volume 10435 of *Lecture Notes in Computer Science*, pages 603–611. Springer International Publishing, 2017.

[272] Dženan Zukić, Aleš Vlasák, Jan Egger, Daniel Hořínek, Christopher Nimsky, and Andreas Kolb. Robust Detection and Segmentation for Diagnosis of Vertebral Diseases Using Routine MR Images. *Computer Graphics Forum*, 33(6):190–204, 2014.

Appendix: Detailed Mammography Meta-Model Results

Table A.1 and Table A.2 provide a detailed overview of model fusion results for the lesion and malignancy prediction, respectively. We report results for the different model types for P_{score} (Random Forest, SVM RBF, MLP) and P_{feat} (CNN) and for varying numbers of considered lesions $n \in \{1, 2, 3, 4, 5\}$. Overall, MLPs outperform random forests and SVMs for prediction score fusion for the lesion as well as the malignancy prediction. MLPs reach a higher AUC score on average (malignancy = 0.772 ± 0.005, lesion = 0.940 ± 0.002) and higher F1 scores as compared to Random Forests and SVMs on test data. When comparing the best P_{score} models (MLPs) with P_{feat}, we observe on average higher measures when performing feature fusion.

			1	2	3	4	5	mean ± std
P_{score}	MLP	AUC	**0.944**	0.940	0.942	0.939	0.937	0.940 ± 0.002
		TPR	0.913	0.927	**0.933**	0.919	0.927	0.924 ± 0.007
		F1	0.929	0.931	0.932	0.929	**0.933**	0.931 ± 0.001
		TNR	**0.829**	0.790	0.771	0.810	0.800	0.800 ± 0.019
P_{score}	Random Forest	AUC	0.928	**0.935**	0.929	0.928	0.923	0.928 ± 0.004
		TPR	0.904	0.913	**0.919**	0.907	0.916	0.912 ± 0.005
		F1	0.919	0.924	0.924	0.923	**0.926**	0.923 ± 0.002
		TNR	0.790	0.790	0.771	**0.810**	0.800	0.792 ± 0.013
P_{score}	SVM RBF	AUC	**0.938**	0.936	0.935	0.929	0.931	0.934 ± 0.003
		TPR	**0.922**	0.910	0.916	0.910	0.913	0.914 ± 0.004
		F1	**0.930**	0.925	0.928	0.919	0.921	0.924 ± 0.004
		TNR	0.800	**0.810**	**0.810**	0.771	0.771	0.792 ± 0.017
P_{feat}	CNN	AUC	0.959	0.950	**0.962***	0.953	0.951	0.955 ± 0.004
		TPR	0.916	0.895	0.956	**0.959***	0.939	0.933 ± 0.024
		F1	0.935	0.922	0.948	**0.952***	0.942	0.940 ± 0.011
		TNR	**0.857***	0.848	0.800	0.819	0.819	0.829 ± 0.021

Table A.1: Detailed evaluation metrics for the *lesion prediction* for different model types and varying number of included lesions n on test data. Bold values indicate the highest values per model and evaluation metric. Best overall values are marked with *.

			1	2	3	4	5	mean ± std
P_{score}	MLP	AUC	0.771	0.772	**0.778**	0.776	0.764	0.772 ± 0.005
		TPR	0.570	0.530	**0.591**	0.544	0.537	0.554 ± 0.023
		F1	**0.601**	0.575	**0.601**	0.574	0.565	0.583 ± 0.015
		TNR	0.837	**0.843**	0.813	0.827	0.820	0.828 ± 0.011
P_{score}	Random Forest	AUC	0.748	0.752	**0.776**	0.767	0.759	0.760 ± 0.010
		TPR	0.577	0.550	0.564	**0.584**	0.550	0.565 ± 0.014
		F1	**0.591**	0.560	0.581	0.582	0.564	0.576 ± 0.012
		TNR	**0.813**	0.793	**0.813**	0.790	0.800	0.802 ± 0.010
P_{score}	SVM RBF	AUC	0.764	0.767	0.763	0.767	**0.768**	0.766 ± 0.002
		TPR	**0.490**	0.477	0.483	0.477	0.477	0.481 ± 0.005
		F1	**0.573**	0.557	0.552	0.544	0.542	0.553 ± 0.011
		TNR	**0.890***	0.883	0.867	0.863	0.860	0.873 ± 0.012
P_{feat}	CNN	AUC	**0.791***	0.775	0.770	0.765	0.776	0.775 ± 0.009
		TPR	0.638	0.436	0.725	**0.698***	0.624	0.624 ± 0.101
		F1	**0.603***	0.518	0.600	0.599	0.565	0.577 ± 0.033
		TNR	0.763	**0.877**	0.657	0.687	0.710	0.739 ± 0.077

Table A.2: Detailed evaluation metrics for the *malignancy prediction* for different model types and varying number of included lesions n on test data. Bold values indicate the highest values per model and evaluation metric. Best overall values are marked with *.