Learning Representations for Medical Image Diagnosis: Impact of Curriculum Training and Architectural Design

Amelia Jiménez Sánchez

DOCTORAL THESIS UPF / 2021

THESIS SUPERVISORS
Prof. Gemma Piella
Prof. Diana Mateus

Dept. of Information and Communication Technologies

Acknowledgement

This work would not have been possible without the support of many lovely and caring people that I feel fortunate to be surrounded by.

First, I would like to thank my co-supervisors Gemma and Diana. Thank you for being so patient and understanding with such a stubborn student. Thank you for giving me freedom to follow my own ideas and for guiding me to grow as a researcher. I am immensely grateful for your support and supervision throughout these years.

I consider myself fortunate to have carried out my PhD thesis thanks to "la Caixa" Foundation at the Department of Information and Communication Technologies (DTIC) of Pompeu Fabra University (UPF). I would like to thank everyone who is/has been part of our research group: BCN MedTech. I would especially like to thank Miguel Ángel González for sharing his experience and providing advice; and Óscar Cámara for being so keen on creating a great working atmosphere with Friday's beers.

Big thanks to my amazing colleagues from the "Plebe" Office: Andrea Urru, Jordi Mill, Mireia Alenyà, Mireia Masias, Pablo Miki and Xabier Morales, without forgetting the ones that left to Barcelona University but loved being with us (and probably miss us): Carlos Martín, Cristian Izquierdo and Víctor Campello. Thanks for sharing many good moments, and of course, micheladas and dinners at "la Pachuca".

The beach volleyball community, that we created, was key to keep mental sanity throughout the PhD. I can't thank you all enough for being part of it: Adrián Martín, Andrea Urru, Arantxa Zapico, Cecília Nunes, Fabio della Valle, Francesco Fabbri, Juan Gómez, Laura Becerra, Lorenzo Porcaro, Miquel Junyent, Pablo Aragón, Rasoul Nikbakht and Silvia Butti.

I feel grateful to have shared the PhD sprint deadline with my submission team: Adrià Arbués, Itziar Zabaleta and Miquel Junyent.

Thanks to everyone in charge of administrative and bureaucratic procedures at the DTIC. Thanks to the wonderful people from secre-

tary, in particular to Vanessa Jiménez, Judith Champion and Ruth Temporal, for always solving any problem. Of course, I am especially and enormously grateful to Lydia García, thank your for easing our PhD journey and for looking after us with so much love.

Other research groups have influenced the work of this thesis:

I would like to thank Coloma Ballester from the Image Processing and Computer Vision group at UPF, who adopted me as a computer vision pupil at the group seminars. Thanks to all the members of the group who provided me with suggestions to improve our work. Special thanks to Patricia Vitoria and Adrià Arbués.

I would like to thank Nassir Navab and Shadi Albarqouni from the Chair for Computer Aided Medical Procedures & Augmented Reality (CAMP) at the Technical University of Munich (TUM). I was lucky to find during my master studies at TUM a great research ecosystem and met many wonderful people. I feel fortunate to have kept contact with them through conferences and visits to Munich. Big thanks for still being there: Antonio Luque, Erini Psallida, Fernando Navarro, Ignacio Sarasúa, Josue Page, Magda Paschali, Nathalia Spier and Santiago Estrada. Our interdisciplinary research would not have been possible without the help and collaboration of the physicians at the Klinikum rechts der Isar. Thanks to Prof. Biberthaler, Dr. Chlodwig Kirchhoff and Dr. Sonja Kirchhoff.

I am grateful to Diana for always being the first one to understand my ideas. I can't thank her enough for her guidance and support. It was great to meet the wonderful and empowered research group that has flourished at the Laboratoire des Sciences du Numérique de Nantes (LS2N) of Ecole Centrale Nantes. Thanks to everyone for being welcoming and kind, especially, thanks to Mickael Tardy, Constance Fourcade and Vanessa González.

Besides research, I would like to thank colleagues from some volunteering initiatives. Special thanks to Magda Paschali for convincing me to apply for the MICCAI Student Board (MSB). It was inspiring to meet undergraduate and graduate students from other research groups across countries, who were so motivated in organizing

and creating events. Thanks to everyone who is/has been part of the MSB. Barcelona was a great ecosystem to participate in dissemination activities. I feel lucky to have been involved with Barcelona Activa (Barcelona City Council), #100tífiques (Catalan Foundation for Research and Innovation), the Inspira STEAM program (Deusto University) and the DTIC Equality Commission at UPF. I would like to thank everyone who has taken part or organized these initiatives. For bridging the gender gap and bringing closer a more egalitarian society, thank you.

There are some people that do deserve a special thank you in this manuscript. Thanks to Daniel for the time that he was by my side. Thanks to Laura Becerra for being brave enough to co-organize with me a beach volleyball tournament (BV DTIC 2018), and to Aurelio Ruiz for his availability, willingness and support to our initiatives. Also, to those that voluntarily came to help and organize a second tournament: Adrián Martín, Cristina González and Pablo Aragón. Overall, big thanks to Hij@s de Caín for the infinite support. You've become indispensable persons in my life. Thanks to Federico Franzoni for bringing calm and welcoming any time at his house. Thanks to Pablo Aragón for making this thesis visually prettier. Thanks to Cecília Nunes for sharing with me the cultural and musical events. Thanks to Rasoul Nikbakht for being so optimistic, always inventing his own path in the mountains, and convincing me that a 48km track in Montseny with 5000m of total elevation gain wasn't a crazy idea for 1 day. Thanks to Juan Gómez for spending so much time listening, bringing perspective to the problems and wonderful musical rhythms from the Pacific. Thanks to Adrián Martín for providing me with a second house, being patient, attentive and having good willingness to organize plans. I don't think anything that I write is going to do justice to the support that I've received from Zaira Pindado, but here are some words. Thanks for being in the good and bad moments, making me feel comfortable talking about everything, and of course, having so much fun together.

Lastly, I would like to finish by thanking my family. These years

would not have been the same without their support, especially, since our life situation got rougher with the outbreak of the COVID-19 pandemic. Enormous thanks to my sister Gloria and my brother Juan, who always have sent me energy, courage and love to keep going. My gratitude goes as well to their couples: Yannick and Giusi; and, of course, to my lovely niece Amélia and my nephew Antoine. These two little ones always bring a smile to my face. I am also grateful to my grandparents: Yaya, Nono, Mami; my aunts: Tita Conchi and Tita Amelia, and Prima Inma for making me feel close despite the distance. Thank you for your everlasting love. I would like to finish dedicating this thesis to my parents, Juan and Antonia, who have been the most loving and supportive persons in the world, having my back and faith in myself at all times. I know you wanted a better future for me and I will be forever grateful for all you have done for me. Os quiero mucho.

Workspace

This work was carried out in the research group Simulation, Imaging and Modelling for Biomedical Systems (SIMBIOSys), part of BCN-MedTech at the Department of Information and Communication Technologies at the Universitat Pompeu Fabra, Barcelona, Spain. This research was carried out in collaboration with the Laboratoire des Sciences du Numérique de Nantes (LS2N) at Centrale Nantes, Nantes, France, and certain collaborations with the Chair for Computer Aided Medical Procedures and Augmented Reality (CAMPAR) at the Technical University of Munich (TUM), Munich, Germany.

Funding

This doctoral project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713673 and by the Spanish

Ministry of Economy [MDM-2015-0502]. Amelia Jiménez-Sánchez has received financial support through the "la Caixa" Foundation (ID Q5850017D), fellowship code: LCF/BQ/IN17/11620013.

Abstract

This thesis investigates two key aspects of learning deep-based image representations for medical diagnosis. The two are confronted with common challenges of medical image databases, namely, the limited number of samples, the presence of unreliable annotations and class-imbalance; as well as, domain shift and data privacy constraints for collaborative learning across institutions. The first part of this thesis concerns the architectural design of deep learning approaches. We explore the importance of localizing the region of interest in the image prior to the classification and the implicit capsule networks' approach to model spatial information. We verify the importance of localization as a preliminary step to the classification, provide a sensitivity analysis of the size of the region of interest, and discuss image retrieval as a clinical use case. We also validate that capsules create equivariance, thus requiring to see fewer viewpoints of the object of interest. The second part of the thesis focuses on easing the optimization of the deep network parameters by gradually increasing the difficulty of the training samples. This gradual increase is based on the concept of curriculum learning and achieved with a data scheduler that controls the order and pace of the samples. We validate the beneficial effect of the curriculum data schedulers in two scenarios. First, we leveraged prior knowledge and uncertainty for the fine-grained classification of proximal femur fractures. In this case, we demonstrated the benefits of our proposed curriculum method under controlled scenarios: with limited amounts of data, under class-imbalance, and in the presence of label noise. Second, we verified the positive effect of the curriculum data scheduler for multi-site breast cancer classification in a federated learning setup.

Resumen

Esta tesis investiga dos aspectos fundamentales del aprendizaje de representaciones profundas de imágenes para el diagnóstico médico. Ambos se enfrentan a los retos comunes de las bases de datos de imágenes médicas, a saber, el número limitado de muestras, la presencia de anotaciones poco fiables y el desequilibrio de clases; así como, la adaptación al dominio ("domain adaptation") y las restricciones de privacidad de datos para el aprendizaje colaborativo entre instituciones. La primera parte de esta tesis se centra en el diseño de arquitecturas para métodos de aprendizaje profundo ("deep learning"). Exploramos la importancia de localizar la región de interés en la imagen antes de la clasificación y el enfoque implícito de redes capsulares ("capsule networks") para modelar la información espacial. Verificamos la importancia de la localización como paso previo a la clasificación, proporcionamos un análisis de sensibilidad del tamaño de la región de interés y discutimos la recuperación de imágenes como caso de uso clínico. También validamos que las cápsulas crean equidistancia, por lo que requieren ver menos puntos de vista del objeto de interés. La segunda parte de la tesis se enfoca en facilitar la optimización de los parámetros de la red aumentando gradualmente la dificultad de las muestras de entrenamiento. Este aumento gradual se basa en el concepto de aprendizaje curricular ("curriculum learning") y se consigue con un programador de datos que controla el orden y el ritmo de las muestras. Validamos el efecto beneficioso de los programadores de datos en dos escenarios. En primer lugar, aprovechamos el conocimiento previo y la incertidumbre para la clasificación granular de las fracturas de fémur proximal. En este caso, demostramos los beneficios de nuestro método basado en aprendizaje curricular bajo escenarios controlados: con cantidades limitadas de datos, desequilibrio de clases y en presencia de anotaciones imprecisas. En segundo lugar, verificamos el efecto positivo del planificador de datos para la clasificación del cáncer de mama en una configuración de aprendizaje federado ("federated learning").

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied. The key for deploying successful machine learning algorithms is on designing the preprocessing pipelines and data transformations that lead to a representation useful for the task. Such feature engineering is important but labor-intensive and requires prior knowledge. For that reason, current machine learning methods have shifted towards automatic feature learning with Deep Learning (DL) and, for the particular case of images, Convolutional Neural Networks (CNNs).

In this thesis, we investigate two key aspects to learn feature representations from medical images for the Computer-Aided Diagnosis (CAD) task. The first aspect is related to the architectural design of the Deep Neural Network (DNN). The second one focuses on easing the optimization of the deep network parameters by gradually increasing the difficulty of the training samples. These two aspects will be confronted by challenges of medical data, namely limited number of annotations or samples, and data privacy for collaborative learning between different hospitals.

1.2 Research goals and context

The overall objective of this thesis is to design methods to improve current automated deep learning systems for computer-aided diagnosis. Medical image datasets, with main focus on the classification task, exhibit some typical characteristics, namely, small amounts of annotated data, class-imbalance and domain shift. These characteristics limit the performance of existing supervised deep learning algorithms. We study these subgoals with different datasets, some of them are public, and three others were collected with medical and industrial collaborators. This overall objective can be broken down into the following contributions:

- Studying alternative architectures, in particular capsule networks, to overcome the limitations of CNNs, especifically the spatial invariance problem. Evaluation of the method for two tasks: mitosis detection and diabetic retinopathy.

- Design of curriculum learning strategies for multi-class classification. We leverage prior or estimated knowledge to design data schedulers. The job of the data scheduler is to determine the order and pace of instances presented to the CNN optimizer. Evaluation on multi-class proximal femur fractures classification.

- Development of a curriculum learning method for a federated setting to improve breast cancer classification. We focus on scheduling the training samples paying special attention to those that are forgotten during the intermediate updates of the global model. The proposed method is combined with unsupervised domain adaptation to deal with domain shift while preserving privacy.

1.3 Outline and contributions

In this dissertation, we investigate deep learning techniques to assist computer-aided diagnosis. The document is divided into two parts: the first one, with two chapters, is focused on architecture and optimization schemes, and the second one, also with two chapters, is focused on training design strategies[1]. The four research chapters are preceded by this introductory chapter and a chapter reviewing the state-of-the-art and background of the thesis. The manuscript then includes a final chapter presenting the conclusions and identifying possible future work directions.

Chapter 2. In this chapter, we review related work on deep learning architectures and training design strategies with a focus on classification. We pay special attention to how these strategies behave under limited amounts of data, imbalance in the class distribution and reliability of the annotations. Finally, we discuss the challenges of employing data from multiple sites and/or devices, as well as the techniques to address them.

PART I: ANALYSIS OF ARCHITECTURE

We first analyze the role of the architecture for different data-challenge scenarios.

Chapter 3 . In this chapter, we investigate the use of CNNs for the localization and fine-grained classification of proximal femur fractures. We demonstrate the importance of the localization of a Region of Interest (ROI) in the X-ray prior to the classification. We provide a sensitivity analysis of the size of the ROI and image retrieval as a clinical use case. We further discuss several strategies of verification of the CNN model for its adoption into daily clinical routine.

[1]The author of this dissertation is the main contributor of the publications listed in this section

The work described in this chapter is included in:

> [JSKA+20] *Amelia Jiménez-Sánchez, Anees Kazi, Shadi Albarqouni, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Sonja Kirchhoff, and Diana Mateus. Precise proximal femur fracture classification for interactive training and surgical planning.* International Journal of Computer Assisted Radiology and Surgery*, 15(5):847–857, April 2020.*

and virtually presented as an oral talk at *the 11th International Conference on Information Processing in Computer-Assisted Interventions — IPCAI 2020.*

Chapter 4 . In this chapter, we studie alternative architectures, in particular capsule networks, to overcome the limitations of CNNs, especifically, the spatial invariance problem. Our method is evaluated for two classification tasks: mitosis detection and diabetic retinopathy. We demonstrate the increased generalization ability of capsule networks *vs.* CNNs when dealing with limited amounts of data and class-imbalance. The performance improvement is a result of the ability of capsule networks to model equivariance, that is, its ability to learn pose parameters along with filter weights. Together with the routing-by-agreement algorithm, this paradigm change requires to see fewer viewpoints of the object of interest, and therefore fewer images, in order to learn the discriminative features to classify them. We found that capsule networks without using data augmentation were able to achieve a similar or better classification performance than CNNs using data augmentation. These results confirm the benefits of equivariance over invariance.

The work described in this chapter is included in:

> [JSAM18] *Amelia Jiménez-Sánchez, Shadi Albarqouni, and Diana Mateus. Capsule networks against medical imaging data challenges. In* Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of

Biomedical Data and Expert Label Synthesis*, pages 150–160, Cham, 2018. Springer International Publishing.*

and presented as a poster an oral talk at the Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (LABELS) at *the 21st International Conference on Medical Image Computing and Computer Assisted Intervention — MICCAI 2018.* The code of this work is publicly available in the following GitHub repository.

PART II: ANALYSIS OF TRAINING DESIGN

In this part, we investigate different manners of optimizing the CNN models with the available training data. We design strategies to control the order, pace and number of images presented to the optimizer. In particular, in the first place, we identify common scheduling elements and unify them into a consolidated curriculum framework. After that, we investigate the use of curriculum into a more challenging scenario of collaborative multi-site learning.

Chapter 5. In this chapter, we design curriculum learning strategies for multi-class classification. Our curriculum learning method is formalized as a data scheduler that determines the order and pace of instances presented to the CNN optimizer. We propose two types of ranking functions to prioritize training data, leveraging: prior knowledge and uncertainty. We validated the benefits of our approach for the classification of proximal femur fractures based on the AO standard, reaching a performance comparable to state-of-the-art and experienced trauma surgeons. Furthermore, in controlled experiments with the MNIST dataset, we show that the proposed method is effective for datasets with class-imbalance, limited or noisy annotations. Results of this chapter are published in:

[JSMK+19] *Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir*

Navab, Miguel A. González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 694–702, Cham, 2019. Springer International Publishing.*

[JSMK+21] *Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. Curriculum learning for improved femur fracture classification: scheduling data with prior knowledge and uncertainty.* Medical Image Analysis (submitted)*, 2021.*

The work of this chapter was presented as a poster at *the 22th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, and later extended. The second manuscript is under review. The code of this work is publicly available in the following GitHub repository.

Chapter 6. In this chapter, we develop a curriculum learning method for a federated learning setting that handles non-IID data from multiple sites. We assume the most challenging scenario, in which each site has mammography systems of different vendors. We show how curriculum learning can boost the performance classification when combined with domain adaptation. Our method is evaluated on high-resolution mammograms from two private and one public dataset for the classification of breast cancer. The resulting manuscript of this chapter is under review:

[JSTB+21] *Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A González Ballester, Diana Mateus, and Gemma Piella. Memory-aware curriculum federated learning for breast*

cancer classification. IEEE Transactions on Medical Imaging (submitted)*, 2021.*

The code of this work is publicly available in the following GitHub repository.

Chapter 7. In this chapter, we conclude this dissertation by summarizing and discussing the main findings and suggesting new directions for future research.

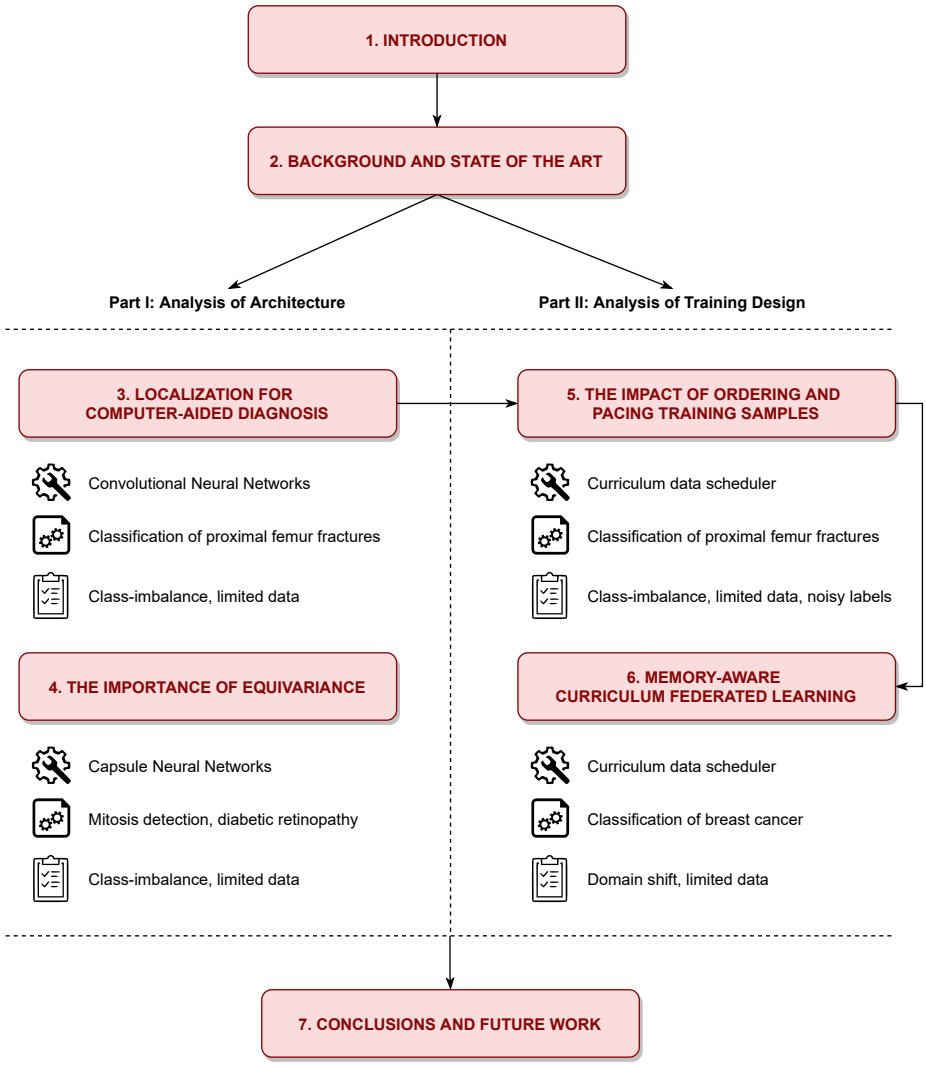

Figure 1.1. Conceptual scheme of this manuscript.

2 Background and State of the Art

2.1 Introduction

In this chapter we provide an overview of DL techniques in medical imaging tasks, with a special focus on CAD and multi-class classification. We pay special attention to works related to training with class-imbalance, limited data, unreliable or uncertain annotations, datasets from multiple sites and suffering from domain shift. This review is intended to make readers familiar with the background of this thesis.

2.2 Data Challenges in Medical Image Analysis

Nowadays, in visual pattern recognition, DL has become the *de facto* standard for classification and detection challenges [Sch15, KSH12a, SZ14, SLJ+15]. The availability of large datasets and the advances in GPU computing power have contributed to the sucess of these techniques. As detailed in Section A.2, CNNs are able to construct hierarchical representations of the input images, which encode com-

plex patterns. However, the sucess of these approaches is limited by the availability of rich and large annotated datasets. Unfortunately, constructing such an ideal dataset for medical image classification presents several challenges. First, given the large patient and disease variability, it is difficult to build a sufficiently large dataset representative of all abnormalities. Furthermore, constructing an annotated dataset requires the expertise and time of physicians, a process which is time consuming and expensive. Second, disease incidence along with the difficulty to collect samples result into class-imbalanced distributions. Even when expert annotations are provided, there might be intra- and/or inter-expert agreement in the patient's diagnosis. Thereby, "noisy" or unreliable annotations are expected in medical datasets. As a consequence, investigating techniques to cope with: restricted amounts of data, class-imbalance, and noisy and limited annotations, without decreasing the performance of the classification task is an area of active research.

To address the above issues, mostly against limited amounts of data, one common-approach has been transfer learning [LKB+17a]. These type of approaches usually compare different networks trained from scratch, *vs.* using off-the-shelf features from the first fully-connected layer [vGSJC15] of a CNN pretrained on ImageNet [TSG+16, SRG+16b] with or without finetuning. In [SRG+16b] different architectures such as AlexNet, GoogLeNet, and VGGNet were investigated for thoraco-abdominal lymph-node detection and interstitial lung-disease classification. It was demonstrated that training from scratch or fine-tuning from ImageNet consistently outperforms applications using off-the-shelf CNN features.

Another approach to cope with limited amounts of data is to extract patches from an image. However, this process often results into the normal class being over-represented and spatial information being of low-quality. For the detection of abnormalities and their further classification, the prior localization of a Region Of Interest (ROI) in the image might be beneficial [dVWdJ+16]. To that end, supervised methods trained with manually-delineated spatial anno-

tations can be leveraged [Gir15]. However, obtaining such ROI annotations is still time consuming and expensive. Therefore, other works [HK16, WPL+17, WLC+19a] have investigated the use of weakly-supervised approaches to improve the classification without a prior detection step. Recently, alternative architectures to CNNs have been proposed. In particular, capsule networks [SFH17] were presented to cope with the CNN's poor modeling of spatial invariance. Capsules encode the *pose* and *presence* of an entity in the image. Their ability to integrate these spatial characteristics makes them attractive to learn representations under the mentioned data challenges.

In the first part of this thesis, we explore the role of architectural design in dealing with spatial information for several classification tasks and under the challenges of medical image datasets. In particular, we explore a supervised localization approach based on imprecise ROI annotations, under different architecture configurations [JSKA+20] (Chapter 3) as well as the implicit capsule's approach to model spatial information [JSAM18] (Chapter 4). In Section 2.3.1, we expose the CNN's shortcomings motivating the later approach.

Instead of localizing an area of interest or enforcing spatial invariance, we can focus on how to select informative images to increase the efficiency of the learning process and reduce the training time. Boosting techniques [Sch03] were proposed to focus the learning process on informative samples. A cascade of learners is created, each trained consecutively, where more emphasis is put on samples misclassified by the previous learners. Then, classification is performed by combining the outputs of each of them. Due to the high computational cost of CNN's optimization, a boosted cascade of them is inefficient. In contrast to boosting techniques, dynamic sampling strategies propose the use of a single learner, which focuses on informative samples during its optimization process. Here, the problem is to define a sampling heuristic for the learner. van Grinsven *et al.* [vGvGH+16] presented a selective data sampling strategy for hemorrhage detection in color fundus images. In particular, they showed that by dynamically increasing the probability of misclassified normal

samples to be selected in each training iteration, classification performance was improved and training was sped-up, verifying the findings in [FSST97, LTY13, EHBG07]. Included in sampling methods, Active Learning (AL) approaches overcome data scarcity by incrementally selecting the most informative unlabeled samples. For instance, Mahapatra *et al.* [MBTR18] presented a framework in which a conditional Generative Adversarial Network (GAN) was employed to generate realistic chest X-ray images with different disease conditions. Then, informative samples were identified using a Bayesian neural network. Related to sampling strategies, Curriculum Learning (CL) rearranges the order in which samples are presented to the optimizer. This order follows the idea of gradual learning, *i.e.* "easy" samples are presented earlier to the optimizer than "hard" ones.

In the second part of this thesis, we investigate CL in the context of medical image classification under dataset constraints. We present a unified CL framework for multi-class classification [JSMK$^+$21] (Chapter 5), and a novel curriculum for collaborative learning [JSTB$^+$21] (Chapter 6). We discuss the state-of-the-art in CL for deep learning (Section 2.4.1). We also review recent literature in Domain Adaptation (DA) (Section 2.4.2) and Federated Learning (FL) (Section 2.4.3) in order to position our most recent work on CL for FL considering DA.

In summary, in this thesis, we deal with image classification tasks under the dataset constraints of medical image analysis, namely, small amounts of labeled data, class-imbalance, domain shift and privacy perservation. In Part I we investigate the role of the architectural design (Chapter 3 and Chapter 4). In Part II we analyze and propose different scheduling approaches based on curriculum learning (Chapter 5 and Chapter 6).

2.3 Architectural Design

In this section, we discuss some limitations of CNNs and the intuition behind capsule networks.

2.3.1 Shortcomings of CNNs

As mentioned earlier, CNNs' success depends on the use of very large databases representative of the full-variability of the source domain. A partial explanation for such requirement comes from two known CNN limitations regarding the spatial invariance and the lack of consideration for the orientation and the spatial relationships between the objects within an image.

An efficient image representation for classification should be able to discard irrelevant information (*e.g.* pose, lighting, etc.). For a CNN to learn such invariant representations, all possible variations of the object should be included in the dataset. CNNs are to some extend translationally invariant. Translational invariance is the ability of a network to detect an object wherever it lies in the image. It is introduced by the pooling layers, which transfer the activation from one layer to the next layer losing spatial information. To handle rotation invariance, data augmentation is a common strategy. Image transformations like rotation, cropping, zoom and others are used to generate variations of the object.

Neurons in a CNN are trained to identify certain shapes or attributes in a image. Initial layers detect low-level features such as edges, circles, etc. Layers further in the network learn higher-level features due to the hierarchical nature of the architecture. However, the output probability will be high when several features of the instance are present in the image. The arrangement of these activations is not relevant for the final output.

2.3.2 Capsule Networks

Capsule Networks [SFH17] were proposed as an alternative feature representation scheme to cope with the above mentioned limitations of classical CNNs, in particular, to cope with their poor modeling of spatial invariance. Capsules are designed to learn the *pose* and *presence* of a capsule's entity, *i.e.* capsules not only identify the existence of a feature, but also its orientation and how it is related to other features. To capture such relationships capsules rely on tensor instead of scalar computations.

Learning in this new capsules regime is achieved through a different optimization scheme: *routing-by-agreement* algorithm, which is designed to maximize agreement between low-level and high-level features. Through dynamic routing, lower level capsules get feedfack from higher-level capsules, and only send their output to the higher level capsules whose output is similar. In this way, the highest-level capsule layers can take into consideration the 'hierarchy of parts' or part-whole relations of an object.

Capsule Layer

A capsule network architecture is composed of only two layers: a first *primary capsule* layer, capturing low-level cues, followed by a specialized *secondary capsule* layer, capable of predicting both the presence and pose of an object in the image. A reconstruction loss and decoder are further included for regularization to boost the coding of the input instantiation parameters. We formally introduce these technical details in Section 4.1.1.

Recently, the use of capsule networks has been investigated for several tasks in medical imaging, such as segmentation [LB18a], image synthesis [HLL+20] and classification [LTB20, MYCVN20].

In this thesis, we study capsules networks to overcome the limitations of CNNs, especifically, the spatial invariance problem. Our method is evaluated for two classification tasks: mitosis detection and diabetic retinopathy. We demonstrate the increased generaliza-

tion ability of capsule networks *vs.* CNNs when dealing with limited amounts of data and class-imbalance. The performance improvement is a result of the ability of capsule networks to model equivariance, that is, its ability to learn pose parameters along with filter weights. Together with the routing-by-agreement algorithm, this paradigm change requires to see fewer viewpoints of the object of interest, and therefore fewer images, in order to learn the discriminative features to classify them. We found that capsule networks without using data augmentation were able to achieve a similar or better classification performance than CNNs using data augmentation. These results confirm the benefits of equivariance over invariance. Details of the proposed method and experimental validation can be found in Sections 4.1.1 and Section 4.2.

2.4 Training Design

In the following we introduce the second type of approaches dealt within the thesis, *i.e.* methods that focus on designing training strategies. Our approaches are derived from the CL concept. Next, we review the state-of-the-art and position our work with respect to other methods in CL, DA and FL. The latter two are important to handle multi-site data suffering from domain shift.

2.4.1 Curriculum Learning

CL is based on the idea that human and animals learn better when information is presented in a meaningful way rather than randomly. A curriculum is an efficient tool for humans to progressively learn from simple concepts to harder tasks. The curriculum breaks down complex knowledge by providing a sequence of learning steps of increasing difficulty. Elman *et al.* [Elm93] brought these ideas from cognitive science to computer science. Bengio *et al.* [BLCW09] made the formal connection of this concept with machine learning using CNNs. In [BLCW09], it was shown that gradually increasing the difficulty

of the task improved both convergence and classification accuracy. Similar to CL, Self-Paced Learning (SPL) [KPK10] also proposes to process the samples in a meaningful order. Different from CL, in SPL the ordering criterion is dynamically estimated from model's performance.

In the following, we review the works that have investigated the use of CL for medical image classification. These works have leveraged: knowledge transfer across tasks [WSMM18, MBN+18], spatial information [JGG+17, PCL+19], data scheduling [YWL+19] and extra information (text or evidence maps) [AEBS+20, ZCCL20]. Some of these works also encountered similar data challenges: limited amounts of data [WSMM18, AEBS+20] and class-imbalance [JGG+17, YWL+19].

Wong *et al.* [WSMM18] exploited the idea of CL to build medical image classifiers with limited data using features from segmentation networks. The proposed CL method outperformed the models pre-trained on ImageNet [DDS+09] and trained from scratch. This CL approach was evaluated for 3D three-class brain tumor MR image classification, and for 2D nine-class segmentation from computed tomography angiography images.

Maicas *et al.* [MBN+18] presented an algorithm that resembles how radiologists are trained. Instead of using all training data and annotations at one, a series of tasks of increasing difficulty composed of smaller datasets was used. The selection of the next task was achieved through a teacher-student curriculum learning [MOCS17], which depends on the model's performance on the tasks and tries to mimic radiologists' training.

To handle extreme class imbalance in the lung nodule detection task, Jesson *et al.* [JGG+17] presented a curriculum adaptive sampling strategy. Their training design is based on CL [BLCW09], starting with patches that only visualize the immediate surrounding of the nodule, and increasing the neighborhood during the optimization process. Showing only nodules would result into an extremely sensitive model with low specificity. Therefore, a schedule is introduced so that the proportion of patches containing nodules to those

that do not is progressively increased, and reaches the data distribution at infinity. The majority of voxels in typical lung images is expected to be predicted as non-nodule. Random sampling leads to a solution that systematically produces false positives. The authors proposed the use of adaptive sampling to favour training examples for which the prediction using recent model parameters produces false results, an instance of hard negative mining [SP98].

Park *et al.* [PCL+19] proposed a two steps CL strategy to detect pulmonary abnormalities in chest-PA X-ray images. On the first step, the CNN was pretrained also using patches around the abnormalities. Then, class activation maps were extracted to provide a weak localization of the abnormalities. All classification metrics were improved with the proposed two steps CL method.

Yang *et al.* [YWL+19] investigated SPL to recognize skin diseases from RGB images while overcoming the existing imbalance scenario. A novel metric termed *complexity of image category* was proposed to integrate both the sample number and the recognition of class difficulty. The intuition is that categories with large number of samples and low intra-class variation are easy to recognize, the goal of Self-Paced Balanced Learning here is to avoid biased results due to the class-imbalance during the learning procedure. Following the formulation of SPL in [MZJ17], the framework schedules the samples according to their complexity and dynamically updates the metric.

Alsharid *et al.* [AEBS+20] proposed a dual-curriculum method for ultrasound image captioning. Here, the method relies on a curriculum built from both image and text information. Their approach leverages textual descriptions that are often rare, and thus lead to small-sized medical datasets. The CL method showed an improvement in all performance metrics for the individual task of image classification as well as for image captioning.

Zhao *et al.* [ZCCL20] leveraged an adaptive dual-curriculum for glaucoma detection under class imbalance and limited amounts of data. Evidence maps were used as training criterion to gradually cure the bias in training data. In particular, the dual curriculum

emphasizes uneven training contributions of data from easy to hard. Their approach significantly improved the convergence speed of the training process and obtained the best classification performances.

Similar to the above previous works, in this thesis we investigate the use of CL to ease the optimization of CNNs for medical image classification. We pay special attention to scenarios with limited or noisy annotations and under class-imbalance. Different from them, we investigate the integration of medical knowledge derived from medical decision trees and inconsistencies in the annotations of multiple experts. We identify the common elements among different data scheduling strategies and present them within a unified framework. In our formulation, we propose two types of ranking functions allowing to prioritize training data: one according to prior-knowledge, and a second one measuring the prediction's uncertainty according to the model's performance (Chapter 5).

To cope with the lack of large rich annotated datasets, new strategies have recently appeared to learn DL methods collaboratively employing data from multiple sites. However, three issues appear in this scenario. The first one concerns the heterogeneity of data from different device systems or hospitals. To cope with such diversity, recent works [PHZS19, LJZ+21] have proposed to integrate Unsupervised Domain Adaptation (UDA) into the FL framework. The second one refers to the preservation of patient privacy and regulations being carefully respected. To address the second challenge, data protection, cryptographic techniques [BIK+17] or differential privacy [DKM+06, DR+14] are employed. The third one is about the requirement of new optimization strategies working on distributed data for collaborative learning. In the FL setting, individual models are trained locally on private data and the central server is responsible for the global aggregation of the local updates. Usually, the communication of the local models to the server occurs a certain number of times every epoch. We introduce a novel curriculum for the FL setup, in which samples that are forgotten after the deployment of the global model are prioritized (Chapter 6). To position our work,

next we discuss DA and FL state-of-the-art.

2.4.2 Domain Adaptation

Machine learning methods are widely used in medical image classification. However, these techniques assume that the training dataset and the test dataset share the same data distribution. When this condition is not satisfied, *i.e.* there is a distribution difference between training and test datasets, the test error generally increases in proportion to the difference. We refer to this problem as *domain shift* [QCSLS09]. This issue is of special importance for multicenter studies. Data coming from multiple sites may be obtained using different devices, scanning image protocols, patient populations, *etc.* We focus on DL techniques to tackle the distribution difference between source and target domains. These techniques have shown their effectiveness for distribution alignment employing maximum mean discrepancy [LCWJ15a], adversarial learning at the feature level [GUA+16, THSD17] or pixel level [BSD+17] or revisiting the batch normalization layer [CPC+17].

In the following, we survey related works on DA for medical image classification. These works perform DA in different manners: employing a zero-bias convolutional autoencoder [AKF+20], using image to image translation [TNA19], revisiting batch normalization layers [WLD20] and leveraging adversarial learning [RHS+18, ZWW+20].

Ahn *et al.* [AKF+20] proposed an UDA approach across different public datasets and problems: medical imaging modality classification, skin disease classification and the detection of multi-drug resistant tuberculosis. The UDA method consisted on a multi-layer zero-bias convolutional autoencoder that constrains the transformation of generic features from a pretrained CNN (for natural images) to non-redundant and locally relevant features for the medical image data. Furthermore, a context-based feature augmentation is added into the scheme to improve the discriminative power of the feature representation.

Tomczak *et al.* [TIM+20] presented an image to image translation DA method for digital staining and classification of leukocytes. A novel combination of image generation with auxiliary tasks such as classification, segmentation and pair-wise reconstruction is introduced. The latter two helped to improve the quality of the generated images.

Wang *et al.* [WLD20] employed domain-specific batch normalization layers [CYS+19]. These layers enable to conduct the feature normalization and estimate internal feature statistics for each site separately. An individual batch normalization layer was assigned to each site independently to tackle statistic discrepancy. Furthermore, to explicitly regularize the latent semantic feature space, the authors proposed to include a contrastive learning [CKNH20] objective. The goal was to encourage robust semantic embeddings that cluster samples regardless of the data source domains. The combination of the domain-specific batch normalization layer and the contrastive learning objective resulted into an improvement in the diagnosis of COVID-19 on two public datasets.

Ren *et al.* [RHS+18] used UDA for classification of prostate histopathology whole-slide images. The adaptation is achieved through adversarial training to find an invariant feature space. A siamese architecture is leveraged to add a regularization on the target domain appropriate for the whole-slide images. In particular, this regularization works on patches within the whole-slide images.

Zhang *et al.* [ZWW+20] presented a collaborative UDA approach to deal with domain shift and label noise. Different target images have diverse discrepancy levels with respect to the source images, therefore the difficulty of the domain alignment is expected to vary for each sample. That is, samples from the target domain that are closer in similarity to the source domain are easier to align than samples that are highly dissimilar. Zhang *et al.*'s domain adversarial approach exploits the classification prediction inconsistency to measure the transferability of source samples. Each inconsistency measure is then used to weigh the domain adversarial loss in the UDA. To

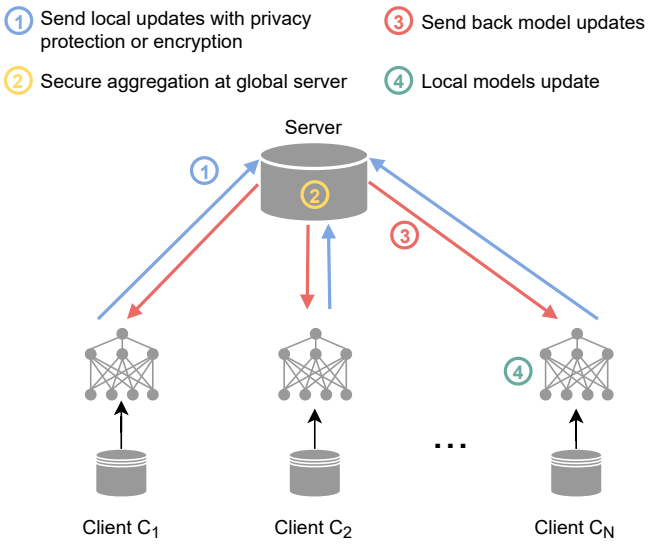

Figure 2.1. Federated learning scenario.

better find hard-to-transfer samples, the diversity of the classifier is maximized via Jensen-Shannon divergence [GPAM+14].

Similar to [ZWW+20], we employ UDA to deal with non-IID data. Different from them, we focus on the classification of breast cancer in a federated setting.

2.4.3 Federated Learning

Acquiring large amounts of labeled medical data requires the effort and expertise of physicians. Therefore, obtaining sufficient data to train DL models is a major challenge. This can be mitigated by the collaboration between institutions. However, sharing medical information among (international) institutions is sensitive in terms of privacy, technical and legal issues in certain regions (GDPR[1] [HM15]

[1]GDPR: EU/UK General Data Protection Regulation

or HIPAA[2] [Ann03]). FL arises from the need of sharing knowledge without sharing sensitive medical data between different hospitals [RHL+20b]. FL enables collaborative and decentralized DL training without sharing any raw patient data [MMR+17a]. The term *federated* was coined because the learning task is solved by a federation of participating models (frequently referred to as *clients*), which are coordinated by a central *server*.

FL frameworks have been formulated in two ways: (i) differential privacy [DKM+06, DR+14], *i.e.* each site training a local model with private data and only sharing model parameters [ZNH+18], (ii) protecting the details of the data using cryptographic techniques [BIK+17], such as secure multi-party computation [MR18] and homomorphic encryption [HHIL+17].

A typical FL scenario is depicted in Fig. 2.1. In this setting, we assume that each local site has data storage and a computing node. Nevertheless, at the global level, only computing is possible. Once that individual models have been trained on private data, there are four key steps in the FL training process: (1) local updates are sent to the global server with privacy protection or encryption, (2) the central server is responsible for the global aggregation of the local updates, (3) the aggregated model parameters are deployed to the local sites, and (4) local models are updated. After that, a new round of local training starts.

Only few FL works have been shown effective on medical images. For instance, for brain tumor segmentation [SRE+18, LMX+19, BWRA21], prediction of disease incidence, patient response to treatment, and other healthcare events [HSQ+19], and lately for classification [GHJK20, AdTBT20, RCS+20, LGD+20, WLD20, YFNA20].

The most common approach for the aggregation of weights in the server is Federated Averaging (`FedAvg`) [MMR+17b]. `FedProx` [LSZ+18] proposes a proximal term to minimize the distance between the local and global models. Alternative aggregation strategies have been proposed in [YFNA20, GHJK20]. Yeganeh *et al.* [YFNA20] proposed in-

[2]HIPAA: Health Insurance Portability and Accountability Act

verse distance aggregation with the objective of handling unbalanced and non-IID data for the classification of dermatoscopic images. Inverse distance aggregation proposes to adaptively weight the contribution of each client by the inverse distance of its parameters to the average model of all clients. The resulting FL model is more robust to noisy and out-of-distribution clients. Grimberg *et al.* [GHJK20] presented a non-linear aggregation FL method, termed *weight erosion*, for the task of survival prediction. It is conceptually related to local fine-tuning, the difference is that a weight erosion scheme is optimized to discard contributions from unhelpful clients as early as possible in the training process.

To deal with heterogeneous image data from multiple sites, device systems or imaging protocols, new techniques have integrated DA into the FL framework. Standard domain approaches require access to source and target data [LCWJ15b, GL15]. However, in the federated setting, data is stored locally and cannot be shared. Recently, federated batch normalization [LJZ+21] and federated adversarial domain adaptation [PHZS19, PKM19] have been proposed to deal with DA under the privacy-preserving requirement. Andreux *et al.* [AdTBT20] proposed a FL approach for the classification of tumorous histopathology image patches from multi-centric data. Local-statistic batch normalization layers are employed to handle DA. The use of these layers results into models that are collaboratively trained, yet site-specific. Li *et al.* [LGD+20] presented a multi-site f-MRI analysis on 1-D signal data using federated adversarial domain adaptation [PHZS19]. Their analysis showed an improvement in classification accuracy and revelead possible brain biomarkers for identifying autism spectrum disorders. Roth *et al.* [RCS+20] investigated breast density classification using FL with data from different institutions. To investigate the effect of domain-shift in the FL method, the authors did not apply any domain adaptation technique to compensate the distribution difference.

Similar to [LGD+20] we employ federated adversarial learning [PHZS19, PKM19] to deal with the alignment between the different

domains. However, unlike Li *et al.* [LGD+20] that analyze 1-D signals extracted from f-MRI, we study the screening of high-resolution mammograms and use CL to boost the classification performance. We validate our strategy on a setup composed of one public and two private clinical datasets with non-IID intensity distributions. Different from [RCS+20], who proposes a FL framework for breast density classification and does not correct the misalignment between the domains, we target the more complex task of breast cancer classification.

2.5 Clinical Applications

The methods proposed in this thesis are evaluated on different medical classification tasks. In particular, we cover: the multi-class classification of proximal femur fractures according to the Arbeitsgemeinschaft für Osteosynthesefragen (AO) standard; detection of diabetic retinopathy, mitosis counting, and breast cancer classification.

We employ three public datasets: TUPAC16 [tup] for mitosis counting, DIARETDB1 [KkKV+] for diabetic retinopathy detection, and INBreast [MAD+12] for the classification of breast lesions. For the latter task, Hera-MI provided two private datasets from different system vendors: Hologic and GE. Institutional board approvals were obtained for each of the datasets (datasets can be shared upon justified requests and subsequent right-holder approvals). The data employed for the classification of proximal femur fractures is from a private dataset. We collected this dataset in collaboration with the trauma surgery department of the Rechts der Isar Hospital in Munich, Germany. The collection of these radiographs was approved by the ethical committee of the Faculty of Medicine from the Technical University of Munich, under the number 409/15 S.

In the following subsections we detail the clinical motivation behind each dataset. Our work aims to provide support to the physicians for each diagnosis task.

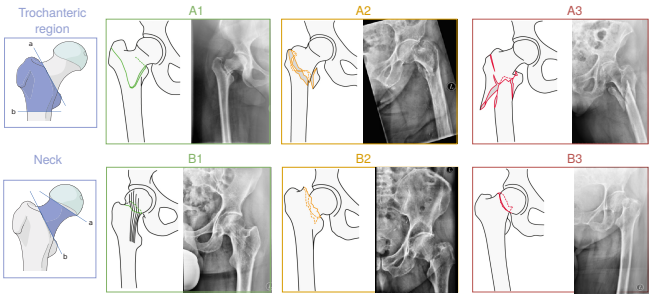

Figure 2.2. AO standard and example X-ray images of proximal femur fractures.

2.5.1 Proximal femur fractures

Proximal femur fractures are a significant cause of morbidity and mortality, giving rise to a notable socioeconomic impact [RYY+15, GCG16]. Elderly population in the western world are especially affected. The incidence of femur fractures increases exponentially from an age of 65 and is almost doubled every five years.

Surgery is the most common and preferred treatment for proximal femur fractures [SSW+15]. The exact classification of the fracture is crucial for deciding the surgical procedure and choosing the surgical implant if needed. The Arbeitsgemeinschaft für Osteosynthesefragen (AO-Foundation) has established a hierarchical classification system for fractures of all bones based on radiographs. For proximal femur fractures, the AO classification has been beneficial, in terms of reproducibility, when compared against other systems such as the Jensen classification [BDE+15]. The AO standard follows a hierarchy according to the location and configuration of the fracture lines, see Fig. 2.2. Fractures of type-A are located in the trochanteric region, and fractures of type-B are those affecting the area around the femur neck. Each type of fracture is further divided into 3 subclasses depending on the morphology and number of fragments of the fracture.

The ability to adequately classify fractures according to the AO standard based on radiographs is acquired through daily clinical rou-

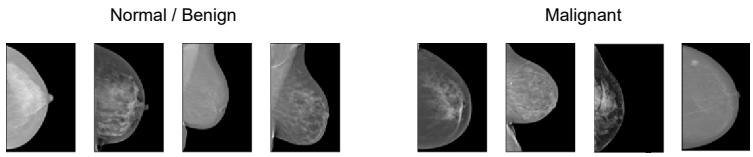

Figure 2.3. Example mammography X-ray images.

tine in the trauma surgery department. Several years are needed until experienced trauma surgeons are significantly differentiated from residents. Inter-reader agreement varies from 66% among residents to 71% among experienced trauma surgeons [Zuc96]. To reach a precise classification, medical students and young trauma surgeons rely on a second opinion to choose the adequate treatment option for the patient.

In Chapter 3 we provide a supervised localization method for the localization of the femur, and target the detection of proximal femur fractures and the classification into {Normal, A, B}. For the more challenging fine-grained classification into {Normal, A1-A3, B1-B3} we present the unified CL framework in Chapter 5. Our approaches are evaluated on a private dataset of about 1000 X-ray images collected at the Rechts der Isar Hospital in Munich (Germany).

2.5.2 Breast cancer

Breast cancer is the most commonly occurring type of cancer worldwide for women [SFS+21]. For this reason, early detection and diagnosis of breast cancer is essential to decrease its associated mortality rate. Mammographic screening has proven to have a positive effect in early diagnosis and has been implemented across many developed countries [DTC+02].

X-ray mammography is currently considered the best imaging method for breast cancer screening and the most effective tool for early detection of this disease [MSMK10]. Among the common findings of abnormalities by radiologists are masses, calcifications, archi-

Table 2.1. Breast Imaging Reporting and Data System Assessment Categories

Category	Description
0	Needs additional imaging evaluation.
1	Negative.
2	Benign finding(s).
3	Probably Benign Finding(s). Follow-up is suggested
4	Suspicious anomaly.
5	Highly suggestive of malignancy.
6	Biopsy proven malignancy.

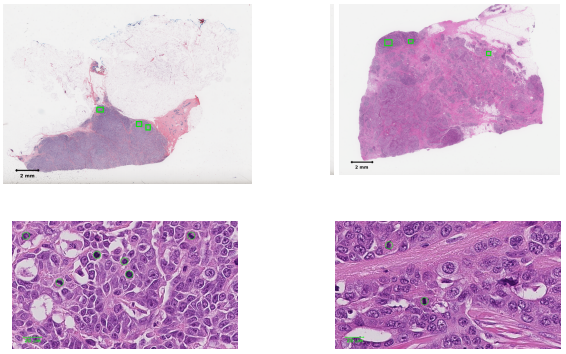

Figure 2.4. Top: Examples of whole-slide images. Bottom: Annotated mitotic. Figure adapted from [tup].

tectural distortion of breast tissue, and asymmetries when comparing the two breasts and the two views [SBWM08]. To standardize the terminology of the mammographic report, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS) scale [D'O96]. This standard includes six categories according to the level of suspicion: category 0, exam is not conclusive; category 1, no findings; category 2, benign findings; category 3, probably benign findings; category 4, suspicious findings; category 5, a high probability of malignancy; and category 6, proved cancer (see Table 2.1). Usually, for categories 4 and 5, biopsy should be considered. The breast composition tissue is also referred by the

ACR, it is an important characteristic related to the breast density shown in X-rays. There are four categories ranging from 1, for low density (fat tissue), to 4, for very high density (dense tissues). The detection of suspicious findings by radiologists is a repetitive and fatiguing task, leading to a 10%-30% rate of undetected lesions [SBWM08, BWY92, KCG+00]. In Chapter 6 we investigate the integration of CL in a federated setting to improve breast cancer classification. Our method is combined with UDA to deal with domain shift while preserving privacy. Our proposed scheme is evaluated on high-resolution mammograms from three different vendors: Siemens (public), GE and Hologic (private).

Tumor proliferation is an important biomarker indicative of the prognosis of breast cancer patients. Patients with high tumor proliferation have worse outcomes compared to patients with low tumor proliferation [vDvdWB04]. The treatment and theratepic plan for the patients depends on the aggressiveness of the tumor. Patients with aggressive tumors are treated with more aggressive therapies, and patients with less aggressive tumors go through more conservative treatments [FPW+00]. Tumor proliferation is assessed on whole-slide images (see Fig. 2.4-top) in a clinical setting by pathologists. This examination is perfomed on histological slides under a miscroscope. The most common method for its quantification is to count mitotic figures (dividing cell nuclei), see 2.4-bottom. Unfortunately, tumor proliferation assessment is a highly subjective and labor-intensive task [VVDJ+16]. In Chapter 4 we compare the performance of CNNs and capsule networks for the mitosis counting task.

2.5.3 Diabetic retinopathy

Diabetic retinopathy affects 1 in 3 adults with diabetes and remains the leading cause of blindness in working-aged adults [WS19]. It is projected to affect 642 million adults by 2040. Thus, it is emerging as a major public health issue worldwide, in particular in low- and middle-income countries. Proper and early treatment of diabetes

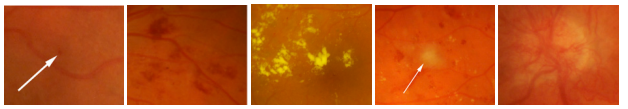

Figure 2.5. Diabetic retinopathy abnormalities: (a) microaneuryms, (b) hemorrhages, (c) hard exudates, (d) soft exudates, (e) neovascularization. Figure courtesy [KKK+07].

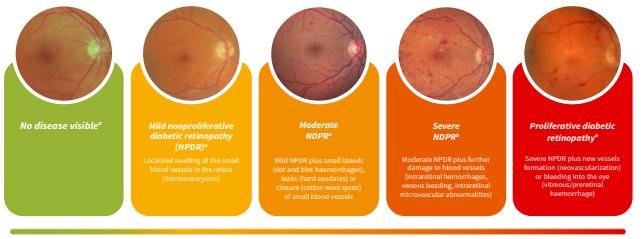

Figure 2.6. Diabetic retinopathy grading. Figure courtesy [O+20].

is cost effective since the implications of poor or late treatment are very expensive. These alarming facts promote the study of automatic diagnosis methods for screening over large populations.

Diabetic retinopathy is a microvascular complication of diabetes, causing abnormalities in the retina, and in the worst case, blindness [KKK+07]. In the initial stage, small changes occur in the retinal capillaries. Abnormalities caused by diabetic retinopathy comprised microaneuryms (Fig. 2.5-a), hemorrhages (Fig. 2.5-b), hard exudates (Fig. 2.5-c), soft exudates (Fig. 2.5-d), and neovascularization (Fig. 2.5-e). Disease severity is depicted in Fig. 2.6. It is graded as mild, moderate or severe non-proliferative diabetetic retinopathy. The first detectable abnormalities are microaneuryms, which are local distensions of the retinal capillary. Their rupture may cause intraretinal hemorrhage. These abnormalities are indicative of mild non-proliferative type. Next, retinal edema and hard exudates appear because of the increased permeability of the capillary walls. At

this stage, the disease is classified as moderate. With time, the obstruction of blood vessels may cause microinfarcts in the retina, which are called soft exudates. When a significant number of intraretinal hemorrhages, soft exudates, or other intraretinal microvascular abnormalities are encountered, the state of the retinopathy is assessed as severe.

Fundus image examination by medical experts is required for the diagnosis of diabetic retinopathy. Furthermore, after it has been diagnosed, regular monitoring is needed due to the progressive nature of the disease. In Chapter 4 we investigate the importance of equivariance for the detection of diabetic retinopathy.

Part I

Analysis of Architecture

3 Localization for Computer-Aided Diagnosis

3.1 Introduction

Proximal femur fractures are a significant problem especially of the elderly population in the western world. Starting at an age of 65 the incidence of femoral fractures increases exponentially and is almost doubled every five years. The consequences of proximal femur fractures have a significant socioeconomic impact since the mortality rate one year after the accident ranges between 14 and 36% [RYY+15, GCG16, BLC14].

In almost all cases, surgical treatment has to be considered the gold standard [BDE+15]. If surgical treatment is decelerated, several complications, as well as an increase in mortality rates, may result [Zuc96, GJR14]. Early detection and classification of proximal femur fractures are crucial for the indication of surgery and, if so, to choose the adequate surgical implant. For the determination of the optimal treatment option, the vascular anatomy of the proximal femur plays an essential role, see Fig. 3.1-(c). These fractures are often described as subcapital, transcervical, or basicervical regarding its location along with displaced *versus* non-displaced. This differ-

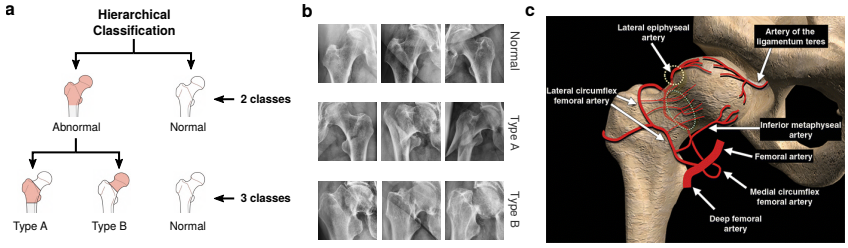

Figure 3.1. (a) Hierarchical classification according to the AO standard. Two scenarios are considered fracture detection (2-class), and classification of the fracture into type A or B (3-class). (b) Examples of regions of interest of X-ray images in our dataset, from top to bottom: healthy femur, fracture type A and B, respectively are shown. (c) Vascular anatomy of the proximal femur, adapted from [SSW+15].

entiation is a key factor because the blood supply to the femoral head is at risk following intraarticular femur fractures. Elderly patients suffer more frequently from transverse subcapital femur fractures arising from rather low-energy trauma compared to younger individuals ($<$ 65 years) showing a tendency to vertical distal femur neck or basicervical fractures resulting from rather high-energy trauma [SWH84, PKP+99]. In this context, the Arbeitsgemeinschaft für Osteosynthesefragen (AO-Foundation) established a generally applicable and valid classification system for fractures of all bones of the skeleton based on X-rays [KMA+18] including the proximal femurs.

In the literature, the AO classification for proximal femur fractures was claimed to present a better reproducibility compared to other classifications such as the Jensen [JDC+05]. In cases of subcapital femur fractures in elderly patients, the Garden classification is more frequently used. However, since the Garden classification describes only subcapital femur fractures, the more extensive AO classification, which includes also intertrochanteric fractures, was used in this study. The AO classification is hierarchical, as shown in Figure 3.1-(a), and is determined by the localization and configurations

of the fracture lines. In case the trochanteric region is broken, fractures are considered as “A" while those affecting the subcapital area “B" fractures. In general, the treatment is oriented to a restoration of mobility and to prevent relapse after surgery. The treatment method depends on the location of the fracture, displacement of fragments, and further concomitant patients' facts like age and functional demands.

The skills for correctly classifying proximal femur fractures are trained during daily clinical routine in the trauma surgery department. However, the learning curve of young trauma surgery residents, especially if working in small peripheral hospitals, is long and shallow. It takes several years of practice to become an expert, as shown by the significant difference in the inter-reader agreement of 66% among residents *vs.* 71% among experienced trauma surgeons [vERMR10]. Currently, young trauma surgeons and medical students majorly rely on the judgment call of colleagues and attendants to achieve a correct classification to choose the best therapeutic option for the patient. Although, there are several online support systems available such as the “bone ninja" or the “AO surgical reference”, these do only demonstrate the different fracture classifications comparable to a textbook. Currently, there is no available automated system capable of classifying X-ray images individually and fracture-specifically.

Therefore, this work aimed to develop a CAD tool based on radiographs to automatically identify proximal femur fractures in a first step, and consecutively classify them according to the AO classification. Such a CAD system can not only help in the correct classification of fractures but also be effective in planning the optimal therapy for the individual patient since the adequate treatment plan arises from the initial classification.

In this work, we show that Convolutional Neural Networks (ConvNets) trained on X-ray images and image-wise class-annotations constitute a suitable predictive model for automatic and on the fly classification of fractures according to the AO standard. We demonstrate the applicability of such models on a clinical dataset of 1347

radiographs. The achieved performance is similar to that of expert radiologists and trauma surgeons reported in [vERMR10]. We further propose a modification of the direct classification workflow considering a localized region of interest (ROI) around the fracture, which further improves the classification results. Finally, we address the question of how to effectively integrate such tool into the clinical routine by

- performing a sensitivity analysis of the size of the localized ROI,
- investigating the potential of retrieval for the training of young trauma surgeons.

3.2 Related Work

From a technical point of view, the automated image analysis of fractures presents significant challenges due to the poor contrast and large variability of the images (see Fig. 3.1-(b)). Such difficulties are exacerbated for proximal femur fractures due to background clutter and the presence of overlapping structures in the pelvic region [WDW+12]. Initial prior work for detection and classification of fractures [AAHR13, BÇ16] focused on conventional machine learning pipelines consisting of preprocessing, feature extraction and classification steps. Predictions are based on hand-crafted features which are sensitive to the low quality of X-ray images. For example, Bayram *et al.* [BÇ16] relied on the number of fragments to classify diaphyseal femur fractures. More recently, deep learning has overcome some limitations of such approaches thanks to the integration of the discriminative feature learning within the predictive models.

The power of ConvNets for fracture detection, that is, for the binary fracture *vs.* not-fractured classification task, has been demonstrated for various anatomical regions, such as spine [RWY+16], wrist [OFM+17], ankle [KCM19], pelvis [WDW+12], and hip [UTG+19]. Badgeley *et al.* [BZOR+19] investigated the complementary added

value of hospital process variables and patient demographics for predicting the presence of fractures comparably to radiographs alone [BZOR+19]. Another studied aspect is the pretraining of the deep models [CHL+19, WLC+19b]. Most works in medical imaging use ImageNet dataset as a pretraining material [UTG+19, BZOR+19, EKN+17]. Instead, Cheng *et al.* [CHL+19] showed that training first the model on an easier task (body part detection on radiographs), resulted in an improvement when later optimizing for the hip fracture detection. Wang *et al.* [WLC+19b] also approached the hip fracture detection employing a sequential pipeline. First, a deeper model was trained to learn high levels of abstraction for binary classification. From this pretrained model, ROIs were extracted in a weak supervised way. Then, a shallower network was trained on the mined ROIs targetting hip and pelvic fracture detection. These methods above point to the effectiveness of ConvNets to assist the radiologist's analysis, reducing the false negative rate and boosting the speed of decisions. However, all of them target still the binary detection problem (abnormal *vs.* not-fractured).

To the best of our knowledge, our team is the first to treat the multi-class classification problem critical for the surgical planning. We demonstrate in this work, that the localization of a ROI is not only important for binary detection, as suggested by Wang *et al.* [WLC+19b], but even more for the multi-class problem. Different from [WLC+19b] and our preliminary work [KAS+17a], where weakly-supervised strategies were explored, here, we focus on the supervised case given its superior performance.

3.3 Methods

Towards improving the clinical training and treatment planning, we aim at developing an automatic CAD system based on ConvNets capable of detecting fractures present on an X-ray image and further predicting its class according to the AO standard. We purposely

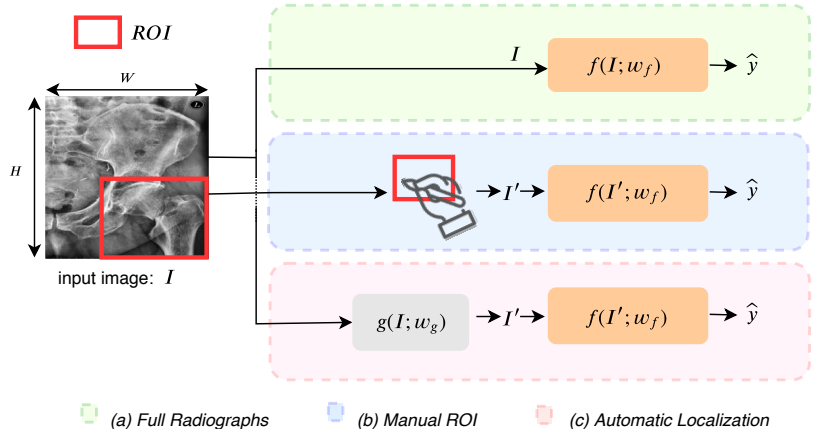

Figure 3.2. Schematic representation of all the considered models. In detail, classification on (a) Full Radiographs, (b) Manual ROIs, and (c) after Automatic Localization.

restrict the choice of architectures and optimization schemes to simple existing methods, and focus instead on developing and evaluating relevant use cases on clinical data.

In practice, given N X-ray images with each image $\mathbf{I} \in \mathbb{R}^{H \times W}$, our goal is to train a *classification model* $f(\cdot)$ that assigns to each image a class label $y \in C$, where C is either $C \subset \{\text{not-fractured, fracture}\}$ for detection or $C \subset \{\text{not-fractured, A, B}\}$ for classification. The conceived model depends on parameters ω_f and provides as output a class prediction $\hat{y} = f(\mathbf{I}; \omega_f)$.

In addition, we define an auxiliary *localization task* $g(\cdot)$ that returns the position $\mathbf{p}$ of the ROI, $\mathbf{I}'$, within the X-ray image such that $\hat{\mathbf{p}} = g(\mathbf{I}; \omega_g)$, where ω_g are the localization model parameters. $\mathbf{p} = \{t_r, t_c, s\}$ is a bounding box of scale s centered at (t_r, t_c). The ROI image $\mathbf{I}' \in \mathbb{R}^{H' \times W'}$ is obtained as $\mathbf{I}' = \mathcal{W}_{\mathbf{p}}(\mathbf{I})$, where $\mathcal{W}_{\mathbf{p}}(\cdot)$ is a warping operator. In the following, we detail three variants of the proposed CAD system to solve and combine the classification and localization tasks differently.

3.3.1 Classification of Full Radiographs

We designed the baseline CAD model to receive a radiograph as input and compute the predicted class in real-time. Figure 3.2-(a) shows a simplified diagram of the used CAD-principle.

Formally, the baseline model is thus $\hat{y} = f(\mathbf{I}; \omega_f)$, where I stands for the whole X-ray image. The mapping f is approximated with a ConvNet optimized to minimize the cross-entropy loss function:

$$\mathcal{L}_{class} = -\sum_{j \in C} y_{j,c} \log(\hat{y}_{j,c}). \tag{3.1}$$

3.3.2 Classification on Manual ROIs

Here, we investigated the influence of localizing a relevant ROI prior to the classification. The ROI was provided by our experts, who manually drew a square containing the head and neck of the femur. We opted for a ROI around the proximal femur instead of a smaller ROI around the fracture, in order to provide contextual information. The cropped image was then used as input to the CAD model.

This second variant is represented in Figure 3.2-(b), and defined as *i.e.* $\hat{y} = f(\mathbf{I}'; \omega_f)$, where $\mathbf{I}'$ denotes the ROI. Independent ConvNets were trained to approximate $f(\cdot)$ in the full image and ROI cases. The later was trained as well with a cross entropy-loss but using a ROI-only dataset.

3.3.3 Classification after Automatic Localization

In this subsection, we focused on an automatic method to localize the ROI within the radiograph. We leveraged the bounding box annotations, manually provided by our experts, to formulate a secondary regression problem aiming to find the ROI in the radiograph. To this end, an auxiliary ConvNet was trained to predict the center and appropriate scale of the bounding box.

We model the localization $g(\cdot)$ and classification $f(\cdot)$ as independent tasks, as depicted in Figure 3.2-(c). The model for classification

is equivalent to the one from the previous section trained on manual ROIs, while $g(\cdot)$ is modeled with a regression ConvNet minimizing the loss:

$$\mathcal{L}_{loc} = \frac{1}{2}\|\mathbf{p} - \hat{\mathbf{p}}\|^2, \tag{3.2}$$

where $\|\cdot\|$ is the ℓ_2-norm, and $\hat{\mathbf{p}}$ is the predicted bounding box. The output localized ROI image is then obtained as $\mathbf{I}' = \mathcal{W}_{\mathbf{p}}(\mathbf{I})$ and fed to $f(\mathbf{I}'; \omega_f)$, only for evaluation.

3.3.4 Model Architectures and Implementation Details

For classification tasks, we used a Residual Network (ResNet-50) [HZRS16a], which was pretrained on ImageNet. The network was trained on radiographs, down-sampled from the original size to 224×224 px. In our case, the categories are the classes in the AO standard (type A and B) and not-fractured. Data augmentation techniques such as translation, scaling and rotation were used. The localization network was designed following AlexNet [KSH12b]. For this architecture, full X-ray images were down-sampled to 227×227 px. All the models were trained on a Linux based workstation equipped with 16GB RAM, Intel(R) Xeon(R) CPU @ 3.50GHz and 64 GB GeForce GTX 1080 graphics card. Stochastic Gradient Descent was used for optimization. All the models were trained until convergence (80 and 200 epochs for classification and localization, respectively). The batch size and momentum were kept constant as 64 and 0.9 for all three models. The learning rate was initialized to 1×10^{-2} for the classification models, and to 1×10^{-8} for the localization network, the decay varied among the different models.

3.4 Experimental Validation

Dataset Collection and Preparation. X-rays of the hip and pelvis of 780 subsequently sampled patients (69% female), with a

Figure 3.3. Localization capabilities of the CAD system. Manually delineated (green) and predicted (blue) bounding boxes for the region of interest in the radiograph.

mean patient age of 75.7 years ± 13.2, diagnosed with proximal femur fractures between 2007 and 2017 at the trauma surgery department of the Rechts der Isar Hospital in Munich were retrospectively gathered in an anonymized manner. The collected images of each patient contained either anterior-posterior (a-p) and lateral view (4%) or only the a-p image. The anterior-posterior views of the pelvis with two hip joints and femora were parted into two, containing one femur each. In most cases, one of the images showed a normal, not-fractured contralateral femur.

Regarding the classification, we looked at two scenarios: fracture detection ("not-fractured" vs. "abnormal") and further division of the "abnormal" class into types "A" and "B". Type "C" fractures were not included in the study as the number of cases was significantly lower than for the other classes. Such fractures are in fact more common in children and follow a different treatment path. For the two-class problem, 780 fracture images and 567 not-fractured images were considered. The same setting was used for the three-class problem considering 327-type A-, 453-type B-fractures and 567-not-fractured X-rays. The dataset was split patient-wise into three parts with the ratio 70%: 10%: 20% to build respectively the training, validation and test set in all presented experiments. To train and test the CAD system, we collected class labels from three clinical experts: one trauma surgeon, one senior radiologist, and one 5th-year resident trauma surgeon (under the supervision of the radiologist).

Each of them evaluated a split of the dataset. The test set was designed to have a class-balanced distribution between classes A, B and not-fractured, consisting of $(55, 60, 55)$ images, respectively. An additional set of 55 not-fractured images was included for the balanced comparison of the two-class scenario (not-fractured: 115 vs. abnormal: 115).

Evaluation Metrics. We used standard classification metrics derived from the confusion matrices: accuracy, precision, recall and F_1-score, and the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve. To evaluate the localization precision of our network, we reported the percentage of ROI centers that are contained in the manually provided bounding boxes. The retrieval task was evaluated using the 11-point precision recall curve [ZZ09].

3.5 Results

3.5.1 Classification on Full Radiographs

Two classification scenarios of proximal femur fractures were evaluated. First, the two-class fracture detection, *i.e.* differentiating between not-fractured and abnormal cases. Second, discriminating among three classes: not-fractured, type A- or type B-fracture. In Table 3.1, we present the accuracy, precision recall and F_1-score for the classification of full X-ray images in two hierarchical scenarios. The performance of the model is maintained when the number of classes was increased from 2 to 3, with an average F_1-score of 84% and 83% respectively.

3.5.2 Classification on Manual ROIs

As it can be seen in Table 1, the use of a region of interest instead of the full image increased all the classification metrics, showing the

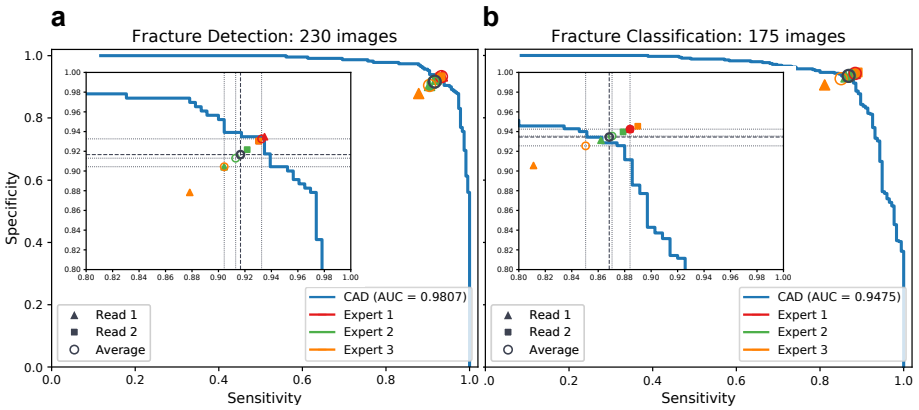

Figure 3.4. Clinical experts and CAD performance. Comparison of specificity measuring the proportion of negatives correctly identified, against sensitivity accounting for the number of positives correctly found, for (a) fracture detection and (b) classification. The set of colors distinguish the CAD system from the individual experts. The filled shapes illustrate the first (triangle) and second (square) readings. The mean performance of every expert is depicted by colored circles, and the average clinical expert by a black circle.

importance of localization as reported in [KAS+17a]. In this case, the network visualizes the characteristics (location, shape, number of fragments) of the fracture at a preferable resolution. F_1-score improvement accounts for 12% (from 0.84 to 0.94) for fracture detection. When the model has to differentiate between fracture type "A" and "B", due to the increased difficulty of the task, the improvement is accounted for 5%. Previous work was reported for only fracture detection, *i.e.* binary classification, our results with an AUC of 0.98 are comparable to state-of-the-art [CHL+19, WLC+19b]. In addition, an AUC of 0.95 was obtained for the three-class problem.

3.5.3 Classification after Automatic Localization

In Figure 3.3, some examples of the manually provided and the predicted bounding boxes are illustrated. Even though there is not a unique way to define a bounding box, we found that the manually defined bounding boxes always (100%) contained the center of the predicted ROIs. Based on the metrics reported in Table 3.1, we observed that the classification model performed similarly on the automatically extracted regions (F_1-score of 93%) and the ones manually provided by the experts (F_1-score of 94%). However, the automatic localization removes the need of an expert intervention during test time.

3.5.4 Expert-level Performance

In order to evaluate the relevance of the previously obtained results, we compared the best-performing model against the individual performance of three clinical experts. We asked three experts (a trauma surgery attendant, a senior radiologist and a trauma surgery attending 1st year) to read the test-set twice with a 2 to 3-week interval between the readings. The evaluation of this set of images took on average 46 minutes. In Table 3.1, the average performance of the experts in the two readings is reported. A ROC analysis was carried

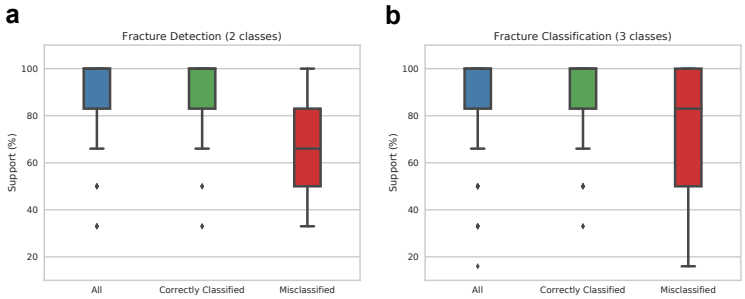

Figure 3.5. Classification robustness and informative disagreement across scales. Percentage of agreeing predictions across different input scales: $[0.75, 1.00, 1.25, 1.50, 1.75, 2.00]$. We gathered the predictions of the scaled regions of interest, and quantified the number of correctly classified for (a) fracture detection and (b) classification. The boxplot shows the median and standard deviation of the support for all test images (in blue), correctly classified (in green) and misclassified (in red).

out by using the ground truth as target classification. To this end, ROC curves were built from the reciprocal relation between sensitivity and specificity calculated for all the possible threshold values. In Figure 3.4, only for visualization purposes, the x-axis has been inverted, *i.e.* we show "Sensitivity" instead of "1-Sensitivity". The average performance between the two readings of each expert compared to the others can be analyzed, but in this case, only one point of the ROC space is obtained. According to the metrics in Table 3.1 and Figure 3.4, our CAD model, trained on manual ROIs, performed similarly compared to the average expert results in fracture classification, and it performed even better regarding the binary fracture detection task.

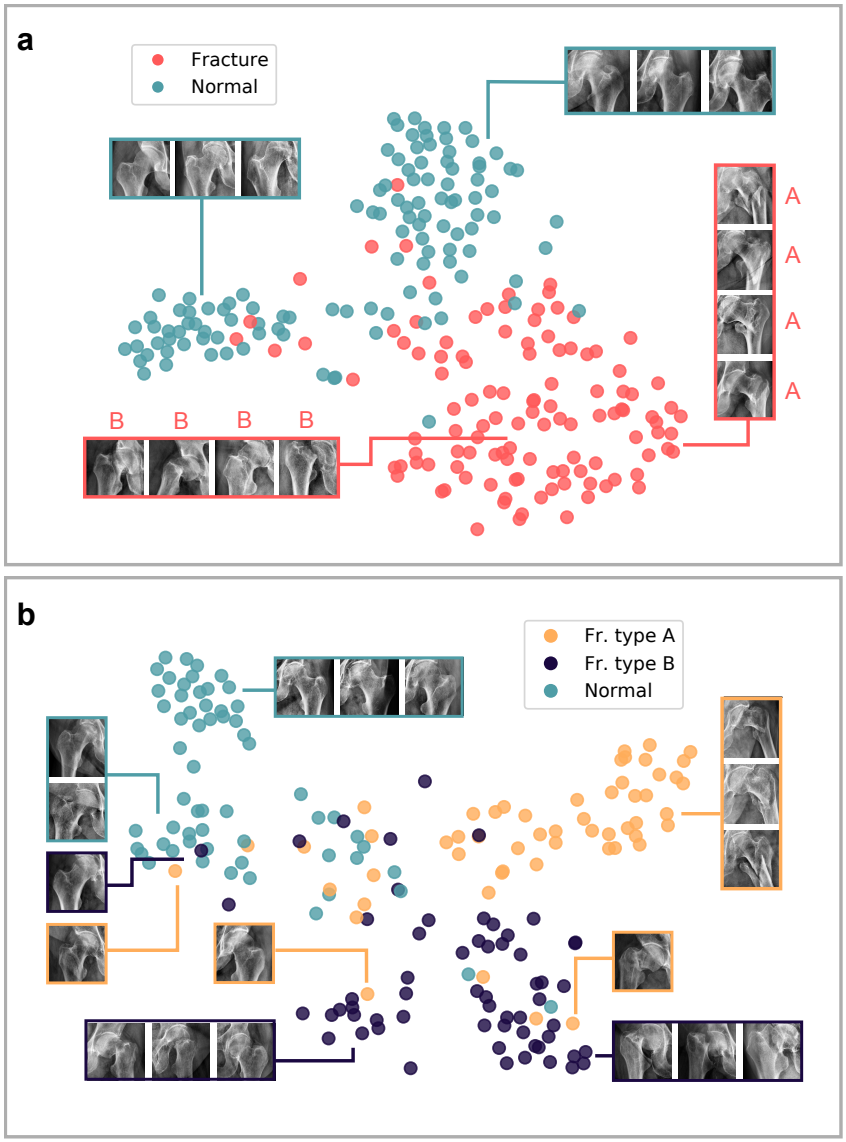

Figure 3.6. Projected 2D space learned by t-SNE. (a) Fracture detection and (b) fracture classification. At the top part, we observe that the model was able to differentiate and group left or right femur. These two clusters were especially differentiated in the not-fractured ("normal") class. Moreover, within the abnormal examples, images of type A and B were differentiated, even if the network was only trained for binary classification.

3.5.5 Robustness and Retrieval

Scale sensitiveness analysis. We further investigated the robustness of our model against the variability of the scale of the predicted ROIs. The predicted bounding boxes were scaled by the following values [0.75,1.00,1.25,1.50,1.75,2.00] and fed to the classification network. We gathered the predictions at each scale and quantified the percentage of correct predictions across scales, i.e. the number of scaled ROIs supporting the correct classification. Results are reported in Figure 3.5. They show a mean support for the correct prediction of 93.82% and 88.35% with 12.34% and 16.58% standard deviation for 2 and 3 classes, respectively. These values demonstrate the robustness and stability of the CAD system to scale variations for most of the cases. Moreover, the disagreement across different scales was shown to be informative of spurious predictions, as suggested by [CP18].

Clinical use case: image retrieval. Here we use the penultimate layer of the network to produce a lower-dimensional representation of each image. We then measure the distance between an unseen query image (from the test set) and the pool of retrievable images (corresponding to the training set), in order to retrieve the most similar cases to the query. We verify the relevance of the retrieval system, by projecting the learned feature representation of the testing images to two dimensions by means of the t-SNE algorithm [MH08]. In the embedded space, depicted in Figure 3.6, the points belonging to different classes for both the 2- and 3-class problems are successfully separated. We evaluate the classification model trained on manual ROIs for retrieval in terms of precision and recall. The precision measures the proportion of relevant images (of the same class as the query) among the retrieved ones. On average, when retrieving 10 images, 9 of the proposed results are relevant. The recall evaluates how many relevant images are retrieved out of the total number of relevant cases. On average, with 100 retrieved images, we recover almost 70% of the relevant cases in the training set. We summarize the results for dif-

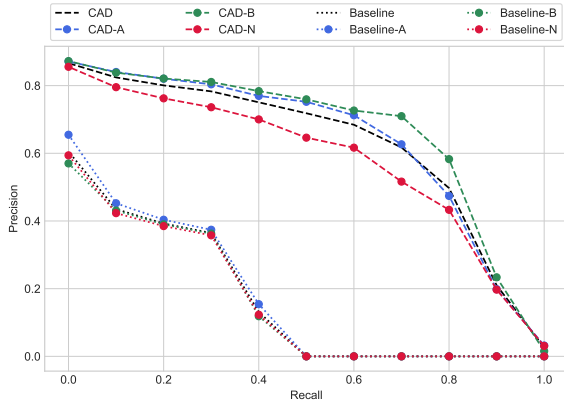

Figure 3.7. Precision *vs.* recall in the image retrieval task. The dashed line represents our best-performing CAD model and the dotted line the baseline. In black color is depicted the average performance, while the colors stand for each of the classes.

ferent numbers of retrieved images $[5, 10, 30, 50, 80, 100, 200, 300, 400]$ in the 11-point precision-recall curve [ZZ09] in Figure 3.7. This curve is based on the Euclidean distances between the query and retrievable images. We compared our CAD model, where the distances are computed on the CNN embedded space, against the distances of the "raw" images (baseline). The proposed CAD retrieval reaches a mean average precision of 0.62 compared to 0.18 of the baseline.

3.6 Discussion

Proximal femur fractures present a huge socioeconomic problem especially in the elderly population. The adequate and exact graduation of these fractures according to the AO classification is highly important for the following treatment and the clinical outcome of the patients. Different to any prior approach, the presented work focuses on a method capable of identifying multiple classes according to a clinical classification standard such as the AO. Our CAD

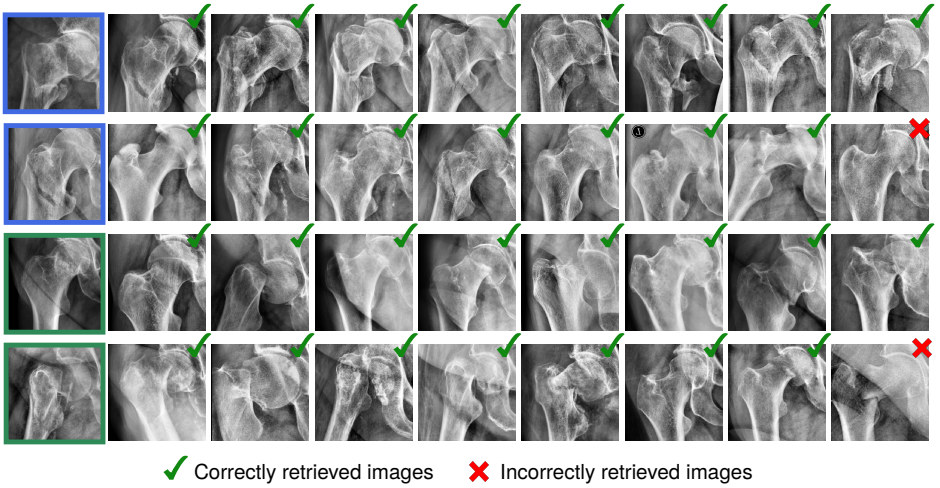

Figure 3.8. Query and retrieval examples. Query images are surrounded by a colored box: A-type fracture (blue) and B-type (green) fracture. For each query, the closest 8 retrieved images are shown. On average, when retrieving 10 images with our CAD model, 9 of the proposed results present the correct classification.

framework exhibits a high F_1-score and AUC of 94% and 0.98 respectively for the two-class problem when differentiating between fracture *vs.* not-fractured, and of 87% and 0.95 for the three-class problem when not-fractured is further divided into type A- and B-fracture. These classification metrics are comparable to state-of-the-art results [CHL+19, WLC+19b]. These high values indicate that our CAD is suitable to be implemented in the daily clinical routing of trauma surgeons treating proximal femur fractures.

Clinical impact. In the short term, our system may assist the trauma surgery residents during their daily clinical training by profiting of a second reading from our CAD tool. In addition, the use of retrieval cases provides an opportunity to focus training through the query of similar and ambiguous cases. Since an inadequate initial

classification may lead to an inappropriate treatment plan, it could support residents in trauma surgery, especially in small peripheral hospitals, to reach a more adequate decision. The most impactful application of such a fracture classification tool would be in the everyday surgery planning, where it is likely that a CAD system could help in reducing fatigue while improving accuracy, given that such a system cannot be affected by bias, experience or workload. In the trauma surgery department, often clinicians are faced with emergency situations and decisions have to be made fast. Due to the fact that our method is able to classify on the fly X-ray images, it could also assist in triaging patients in the emergency room. In cases of proximal femur fractures, time until diagnosis is critical. In fact early surgery and mobilization have been identified as key factors in reducing the number of complications after surgery and mortality [FWD+18]. In treating fracture neck femurs (B-type fractures), classification remains the best option to reduce the risk of complications like non-union and avascular necrosis in treating fracture neck femurs [MB12]. In this context, a waiting time of 24 hours until surgery have been shown to be associated with a greater risk of 30-day mortality [PRW+17].

Adoption into clinical practice. In order to favor the integration of the proposed CAD system into the daily clinical routine, we propose three layers of verification. First, along with the predicted class labels, we provide as an additional output the localization of the fracture in the form of a bounding box. A proper working system provides a bounding box leaving out any irrelevant regions. If it were the case that the system predicts an area of support outside the expected anatomy, a supportive minimal interaction tool is procured, consisting of two clicks to manually select the ROI around the fracture. Classification on the selected ROI leads to both a positive and significant improvement in the classification as described in Results section. A second verification is the agreement of the class predictions over several scales (see Figure 3.5), where we found that

disagreement could point out misclassification examples. Finally, we use the learned feature space of the network to retrieve similar images (see Figure 3.8). Such a retrieval system may be used by residents to learn variants of a single fracture type or even useful for experts to analyze complex cases in comparison with the retrieved samples. In this way, our method helps to speed up the diagnosis and treatment planning for complex fractures cases.

Technical limitations. From the technical side, our dataset suffers from a high imbalance in the distribution of the classes when considering the subtypes of A- and B- fractures. Such scenarios require more complex (and thus less interpretable models) [KAS+17a] or different optimization strategies [BLCW09, JSMK+19]. We further plan to investigate weighting schemes, for example based on the uncertainty of the model [GG16], or relying on triplet metric learning [HLCLT16]. Exploring bootstrapping strategies could be one way of handling noisy labels [RLA+15]. One could also consider modeling label's uncertainty to estimate the uncertainty due to intra or inter-observer variability [TNA19].

3.7 Conclusions

We have proposed a CAD scheme for the detection and further classification of proximal femur fractures achieving results comparable to state-of-the-art performance for the binary fracture detection task. Moreover, we show for the first time, an in-depth evaluation of an automatic system for the multi-class problem according to the AO system. This level of categorization is crucial for planning the treatment either conservatively or surgically, and if so, to choose the adequate surgical implant. The localization of the region of interest was highly accurate, all the predicted centers of the ROI were contained in the original bounding box. The sensitivity of the system to the size of the ROI was analyzed in detail; we found that disagreement

in classification at different ROI sizes could signal the potential for misclassification. We presented a clinical use case of retrieval to assist the training of trauma surgery residents, especially for those working in small peripheral hospitals. Finally, we discussed several strategies of verification to favor the adoption of our CAD tool into the daily clinical routine.

	2 classes	3 classes			
	Abnormal	Type A	Type B	Normal	Avg.
Full Radiographs					
Accuracy	83%	86%	87%	**94%**	89%
Precision	78%	86%	78%	88%	84%
Recall	83%	67%	83%	95%	82%
F_1-score	84%	76%	82%	**91%**	83%
Manual ROIs					
Accuracy	**93%**	**91%**	**91%**	91%	**91%**
Precision	93%	**98%**	**87%**	81%	**88%**
Recall	**94%**	**75%**	88%	**97%**	**87%**
F_1-score	**94%**	**85%**	**88%**	88%	**87%**
Automatic Localization					
Accuracy	**93%**	89%	87%	**94%**	90%
Precision	**94%**	90%	77%	90%	86%
Recall	93%	73%	**90%**	92%	85%
F_1-score	93%	81%	83%	**91%**	85%
Clinical Expert					
Accuracy	92%	92%	89%	93%	**91%**
Precision	92%	90%	79%	94%	**88%**
Recall	92%	83%	92%	86%	**87%**
F_1-score	92%	86%	95%	90%	**87%**

Table 3.1. Classification metrics for evaluation of the three compared methods: full radiographs, manually defined ROIs (Manual ROIs), and after automatically predicting the ROIs (Automatic Localization); and the average clinical expert. Accuracy, precision, recall and F_1-score of our models. The highest metric values across the three models are highlighted in bold for each metric and classification type.

4 The Importance of Equivariance

4.1 Introduction

Currently, numerous state-of-the-art solutions for medical image analysis tasks such as computer-aided detection or diagnosis rely on Convolutional Neural Networks (ConvNets) [LKB+17b]. The popularity of ConvNets relies on their capability to learn meaningful and hierarchical image representations directly from examples, resulting in a feature extraction approach that is flexible, general and capable of encoding complex patterns. However, their success depends on the availability of very-large databases representative of the full-variations of the input source. This is a problem when dealing with medical images as their collection and labeling are confronted with both data privacy issues and the need for time-consuming expert annotations. Furthermore, we have poor control of the class distributions in medical databases, *i.e.* there is often an imbalance problem. Although strategies like transfer learning [ZLZ+17a], data augmentation [VV17b] or crowdsourcing [ABA+16] have been proposed, data collection and annotations is for many medical applications still a bottleneck [CAL+17].

ConvNets' requirement for big amounts of data is commonly jus-

tified by a large number of network parameters to train under a non-convex optimization scheme. We argue, however, that part of these data requirements is there to cope with their poor modeling of spatial invariance. As it is known, purely convolutional networks are not natively spatially invariant. Instead, they rely on pooling layers to achieve translation invariance, and on data-augmentation to handle rotation invariance. With pooling, the convolution filters learn the distinctive features of the object of interest irrespective of their location. Thereby losing the spatial relationship among features which might be essential to determine their class (e.g. the presence of plane parts in an image does not ensure that it contains a plane).

Recently, capsule networks [SFH17] were introduced as an alternative deep learning architecture and training approach to model the spatial/viewpoint variability of an object in the image. Inspired by computer graphics, capsule networks not only learn good weights for feature extraction and image classification but also learn how to infer pose parameters from the image. Poses are modeled as multidimensional vectors whose entries parametrize spatial variations such as rotation, thickness, skewness, *etc.* As an example, a capsule network learns to determine whether a plane is in the image, but also if the plane is located to the left or right or if it is rotated. This is known as *equivariance* and it is a property of human one-shot learning type of vision.

In this work, we experimentally demonstrate that the equivariance properties of CapsNets reduce the strong data requirements, and are therefore very promising for medical image analysis. Focusing on computer-aided diagnosis (classification) tasks, we address the problems of the limited amount of annotated data and imbalance of class distributions. To ensure the validity of our claims, we perform a large number of controlled experiments on two vision (MNIST and Fashion-MNIST) and two medical datasets that targets: mitosis detection (TUPAC16) and diabetic retinopathy detection (DIARETDB1). To the best of our knowledge, this is the first study to address data challenges in the medical image analysis community

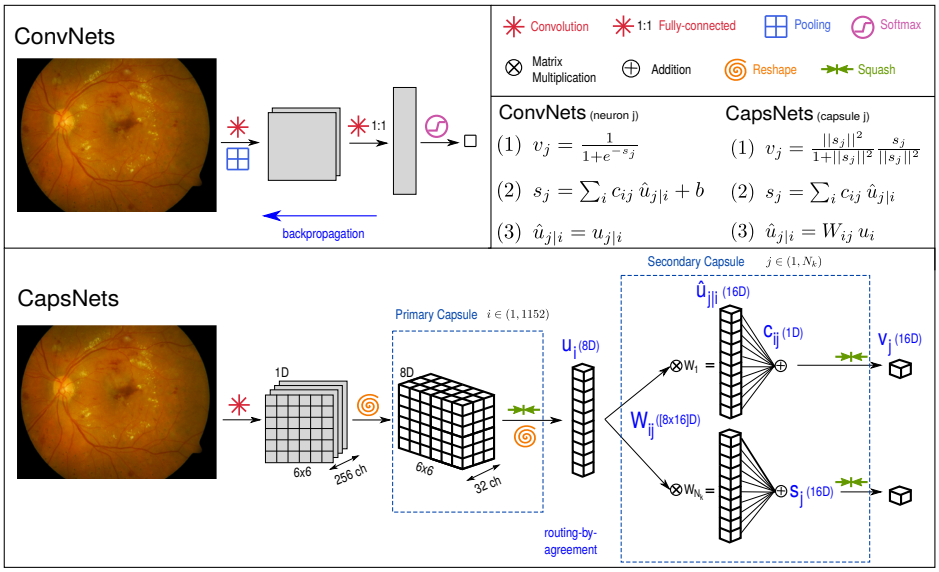

Figure 4.1. Comparison of the flow and connections of ConvNets *vs.* CapsNets. Eq. (1) shows the difference between the sigmoid and squashing functions. Eq. (2) is a weighted sum of the inputs (ConvNets use bias). In CapsNets, c_{ij} are the coupling coefficients. In (3), $\hat{u}_{j|i}$ is the transformed input to the *j-th* capsule/neuron. In CapsNets, the input from the *i-th* capsule is transformed with the weights W_{ij}. While in ConvNets, the raw input from the previous neuron is used.

with Capsule Networks.

In the following, we focus on the image classification problem characteristic of computer-aided diagnosis systems. Our objective is to study the behavior of Capsule Networks (CapsNets) [SFH17] in comparison to standard Convolutional Networks (ConvNets) under typical constraints of biomedical image databases, such as a limited amount of labeled data and class imbalance. We discuss the technical advantages that make CapsNets better suited to deal with the above-mentioned challenges and experimentally demonstrate their improved performance.

4.1.1 Capsule vs Convolutional Networks

Similar to ConvNet approaches, CapsNets build a hierarchical image representation by passing an image through multiple layers of the network. However, as opposed to the tendency towards deeper models, the original CapsNet is formed with only two layers: a first *primary caps* layer, capturing low-level cues, followed by a specialized *secondary caps*, capable of predicting both the presence and *pose* of an object in the image. The main technical differences of CapsNets w.r.t. ConvNets are:

i) Convolutions are only performed as the first operation of the *primary caps* layer, leading as usual to a series of *feature channels*.

ii) Instead of applying a non-linearity to the scalar outputs of the convolution filters, CapsNets build tensors by grouping multiple feature channels (see the grid in Fig. 4.1). The non-linearity, a *squashing* function, becomes also a multidimensional operation, that takes the $j-th$ vector s_j and restricts its range to the [0,1] interval to model probabilities while preserving the vector orientation. The result of the squashing function is a vector v_j, whose magnitude can be then interpreted as the probability of the presence of a capsule's entity, while the direction encodes its pose. v_j is then the output of the capsule j.

iii) The weights W_{ij} connecting the i primary capsule to the the $j-th$ secondary capsule are an affine transformation. These transformations allow learning part/whole relationships, instead of detecting independent features by filtering at different scales portions of the image.

iv) The transformation weights W_{ij} are not optimized with the regular backpropagation but with a *routing-by-agreement* algorithm. The principal idea of the algorithm is that a lower level capsule will send its input to the higher level capsule that *agrees* better with its input, this way is possible to establish the connection between lower- and higher-level information (refer to [SFH17] for details).

v) Finally, the output of a ConvNet is typically a softmax layer

with cross-entropy loss: $\mathcal{L}_{ce} = -\sum_x g_l(x)\ log(p_l(x))$.

Instead, for every secondary capsule, CapsNet computes the margin loss for class k:

$$\mathcal{L}_k = T_k\ \max(0, m^+ - ||\mathbf{v}_k||)^2 + \lambda\ (1 - T_k)\ \max(0, ||\mathbf{v}_k|| - m^-)^2, \quad (4.1)$$

where the one-hot encoded labels T_k are 1 iff an entity of class k is present and $m^+ = 0.9$ and $m^- = 0.1$, i.e. if an entity of class k is present, its probability is expected to be above 0.9 ($||\mathbf{v}_k|| > 0.9$), and if it is absent $||\mathbf{v}_k|| < 0.1$. Since the threshold is not set as 0.5, the marginal loss forces the distances of the positive instances to be close to each other, resulting in a more robust classifier. The weight $\lambda = 0.5$.

As regularization method, CapsNet uses a decoder branch composed of two fully connected layers of 512 and 1024 filters respectively. The loss of this branch is the mean square error between the input image x and its reconstruction $\hat{x}$ both of size $N \times M$,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N \cdot M} \sum_{n=1}^{N} \sum_{m=1}^{M} (x(n,m) - \hat{x}(n,m))^2) \quad (4.2)$$

The final loss, is a weighted average of the margin loss and the reconstruction loss $\mathcal{L}_{total} = \sum_{k=1}^{N_k} \mathcal{L}_k + \alpha\ \mathcal{L}_{MSE}$.

4.1.2 Medical Data Challenges

It is frequent for medical image datasets to be small and highly imbalanced. Particularly, for rare disorders or volumetric segmentation, healthy samples are the majority against the abnormal ones. The cost of miss-predictions in the minority class is higher than in the majority one since high-risk patients tend to be in the minority class. There are two common strategies to cope with such scenarios: i) increase the number of data samples and balance the class distribution, and ii) use weights to penalize stronger miss-predictions of the minority class.

We propose here to rely on the equivariance property of CapsNets to exploit the structural redundancy in the images and thereby reduce the number of images needed for training. For example, in Fig. 4.1, we can see a fundus image in which diabetic retinopathy is present. There are different patterns present in the image that could lead to a positive diagnosis. Particularly, one can find soft and hard exudates or hemorrhages. While a ConvNet would tend to detect the presence of any of these features to make a decision, CapsNet routing algorithm is instead designed to learn to find relations between features. Redundant features are collected by the routing algorithm instead of replicated in several parts of the network to cope with invariance. We claim that the above advantages directly affect the number of data samples needed to train the networks. To demonstrate our hypothesis we have carefully designed a systematic and large set of experiments comparing a traditional ConvNet: LeNet [LBBH98] and a standard ConvNet: Baseline from [SFH17], against a Capsule Network [SFH17]. We focus on comparing their performance with regard to the medical data challenges to answer the following questions:

- How do networks behave under decreasing amounts of training data?
- Is there a change in their response to class-imbalance?
- Is there any benefit from data augmentation as a complementary strategy?

To study the generalization of our claims, our designed experiments are evaluated on four publicly available datasets for two vision and two medical applications: i) Handwritten Digit Recognition (MNIST), ii) Clothes Classification (FASHION MNIST), iii) Mitosis detection, a sub-task of mitosis counting, which is the standard way of assessing tumor proliferation in breast cancer images (TUPAC16 challenge [tup]), and iv) Diabetic Retinopathy, an eye disease, that due to diabetes could end up in eye blindness over time. It is detected

	Conv1	Pool1	Conv2	Pool2	Conv3	-	FC1	Drop	FC2
LeNet	5×5 6 ch	2×2	5×5 16 ch	2×2	✗	-	1×1 120 ch	✗	1×1 84 ch
Baseline	5×5 256 ch	✗	5×5 256 ch	✗	5×5 128 ch	-	1×1 328 ch	✓	1×1 192 ch
	Conv1	Pool1	Conv2	Pool2	Caps1	Caps2	FC1	Drop	FC2
CapsNet	9×9 256 ch	✗	9×9 256 ch	✗	1152 caps 8D	N_k caps 16D	1×1 512 ch	✗	1×1 1024 ch

Table 4.1. Details of each of the architectures. For convolution, we specify the size of the kernel and the number of output channels. In the case of pooling, the size of the kernel. And for capsule layers, first, the number of capsules and, in the second row, the number of dimensions of each capsule.

by a retinal screening test (DIARETDB1 dataset). Next, we provide some implementation details of the compared methods.

Architectures. Since research of capsules is still in its infancy, we pick the first ConvNet, LeNet [LBBH98] for a comparison. Though this network has not many parameters (approx. 60K), it is important to notice the presence of pooling layers which reduce the number of parameters and lose the spatial relationship among features. For a fairer comparison, we pick another ConvNet with similar complexity to CapsNet, in terms of training time, that has no pooling layers, which we name hereafter Baseline and was also used for comparison in [SFH17].

LeNet has two convolutional layers of 6 and 16 filters. Kernels are of size 5x5 and stride 1. Both are followed by a ReLU and pooling of size 2x2. Next, there are two fully connected layers with 120 and 84 filters. **Baseline** is composed of three convolutional layers of 256, 256, 128 channels, with 5x5 kernel and stride of 1. Followed by two fully connected layers of size 382, 192 and dropout. In both cases, the last layer is connected to a softmax layer with cross-entropy loss. For **CapsNet [SFH17]**, we consider two convolutional layers of 256 filters with kernel size of 9x9 and stride of 1. Followed by two capsule layers of 8 and 16 dimensions, respectively, as depicted in Fig. 4.1. For

each of the 16-dimensional vectors that we have per class, we compute the margin loss like [SFH17] and attach a decoder to reconstruct the input image. Details are summarized in Table 4.1.

Implementation. The networks were trained on a Linux-based system, with 32 GB RAM, Intel(R) Core(TM) CPU @ 3.70 GHz and 32 GB GeForce GTX 1080 graphics card. All models were implemented using Google's Machine Learning library TensorFlow [1]. The convolutional layers are initialized with Xavier weights [GB10]. All the models were trained in an end to end fashion, with Adam optimization algorithm [KB14], using grayscale images of size 28×28. The batch size was set to 128. For MNIST and Fashion-MNIST, we use the same learning rate and weight for the reconstruction loss as [SFH17], while for AMIDA and DIARETDB1 we reduced both by 10. If not otherwise stated, the models were trained for 50 epochs. The reported results were tested at minimum validation loss.

4.2 Experimental Validation

Our systematic experimental validation compares the performance of LeNet, a Baseline ConvNet and CapsNet with regard to the three mentioned data-challenges, namely the limited amount of training data, the class-imbalance, and the utility of data-augmentation. We trained in total 432 networks, using 3 different architectures, under 9 different data conditions, for 4 repetitions, and for 4 publicly available datasets. The two first datasets are the well known MNIST [LC10] and Fashion-MNIST [XRV17], with 10 classes and, 60K and 10K images for training and test respectively.

For *mitosis detection*, we use the histological images of the first auxiliary dataset from the TUPAC16 challenge [tup]. There are a total of 73 breast cancer images, of $2K \times 2K$ pixels each, and with the annotated location coordinates of the mitotic figures. Images are

[1] https://www.tensorflow.org/

normalized using color deconvolution [VPS+16] and only the hematoxylin channel is kept. We extract patches of size 100×100 pixels that are downsampled to 28×28, leading to about 60K and 8K images for training and test respectively. The two classes are approximately class-wise balanced after sampling.

For the *diabetic retinopathy detection*, we consider DIARETDB1 dataset [KkKV+]. It consists of 89 color fundus images of size $1.1K \times 1.5K$ pixels, of which 84 contain at least mild signs of the diabetic retinopathy, and 5 are considered as normal. Ground truth is provided as masks. We enhance the contrast of the fundus images by applying contrast limited adaptive histogram equalization (CLAHE) on the lab color space and keep only the green channel. We extract patches of 200×200 pixels that are resized to 28×28. This results in about 50K and 3K images for training and test respectively. They are approximately class-wise balanced after sampling.

4.2.1 Limited amount of training data

We compare the performance of the two networks for the different classification tasks when the original amount of training data is reduced to 50%, 10%, 5%, and 1% while keeping the original class distribution. We run each of the models for the same number of iterations that are required to train 50 full epochs using all the training data. Early-stop is applied if the validation loss does not improve in the last 20 epochs.

The results are shown in Table 4.2. For almost all scenarios CapsNet performs better than LeNet and Baseline. We can observe in Figure 4.2 how for MNIST the gap is higher for a small amount of data and is reduced when more data is included. LeNet with 5% of the data has a similar performance to CapsNet, and better than Baseline, with 1% of the data for DIARETDB1. We attribute this behavior to the structures that are present in this type of images. All the experiments validated the significance test with a p-value $<$ 0.05, except for those on the TUPAC16 dataset, we presume this is

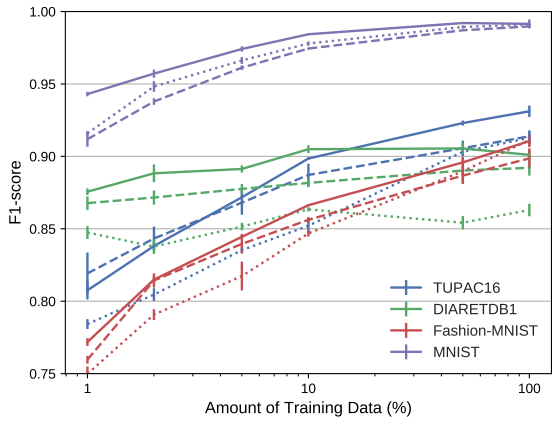

Figure 4.2. Mean F_1-score and standard deviation (4 runs) for different amounts of training data. Solid line: CapsNet, dotted line: Baseline, and dashed line: LeNet.

associated to the CapsNet limitations that we present in Section 6.7.

Table 4.2. F_1-scores using different amounts of training data.

Training Data	1%			5%			10%			50%		
	LeNet	Base.	CapsNet	LeNet	Base.	CapsNet	LeNet	Base.	CapsNet	LeNet	Base.	CapsNet
TUPAC16	**0.822**	0.784	0.809	0.872	0.835	0.872	0.890	0.852	**0.898**	0.908	0.903	**0.923**
DIARETDB1	0.870	0.847	**0.875**	0.877	0.852	**0.893**	0.883	0.863	**0.907**	0.895	0.854	**0.908**
Fashion-M.	0.759	0.749	**0.772**	0.841	0.817	**0.846**	0.856	0.847	**0.866**	0.885	0.889	**0.896**
MNIST	0.909	0.916	**0.943**	0.961	0.966	**0.975**	0.975	0.978	**0.985**	0.987	0.989	**0.992**

4.2.2 Class-imbalance

For the medical datasets, we simulate class imbalance by reducing to 20% one of the two classes. Initially, we reduce abnormal class and, afterward, the healthy class. For the other two datasets, we decrease two classes at the same time. For MNIST, we first consider reducing the classes "0" and "1" and secondly, the classes "2" and "8". Similar for Fashion-MNIST, we reduce the classes "T-shirt/top" and "Trouser", and in the second scenario, "Pullover" and "Shirt".

Table 4.3. F_1-scores for different class-imbalance scenarios.

Scenario	Balanced			Imbalanced 1			Imbalanced 2		
	LeNet	Base.	Caps.	LeNet	Base.	Caps.	LeNet	Base.	Caps.
TUPAC16	0.914	0.913	**0.932**	0.881	0.813	**0.892**	0.905	0.874	**0.909**
DIARETDB1	0.895	0.863	**0.899**	0.869	0.839	**0.887**	0.889	0.874	**0.898**
Fashion-M.	0.899	**0.911**	0.910	0.890	**0.902**	0.889	0.871	**0.881**	0.863
MNIST	0.989	0.991	0.991	0.988	0.989	**0.993**	0.985	0.987	**0.992**

Table 4.4. Mean F_1-score with and without data augmentation.

Data Augmentation	No			Yes		
	LeNet	Base.	Caps.	LeNet	Base.	Caps.
TUPAC16	0.904	0.892	**0.914**	0.914	0.913	**0.932**
DIARETDB1	0.883	0.864	**0.895**	0.892	0.863	**0.899**
Fashion-MNIST	0.899	**0.911**	0.910	0.902	0.911	**0.913**
MNIST	0.989	0.991	0.991	0.990	0.993	**0.994**

In Table 4.3 results are reported. Again, CapsNet surpasses the performance of ConvNets for all cases, except for Fashion-MNIST where the f1-scores are similar. At least one of the imbalance cases verified the significance test for all datasets.

4.2.3 Data augmentation

In the last series of experiments, we compare the performance of the three networks using data augmentation, a common technique to increase the amount of training data and balance class distributions. The original dataset is augmented with ±10 degrees rotations, with a translation of ±30 pixels for medical datasets, and with flips (horizontal for Fashion-MNIST and, both horizontal and vertical for TUPAC16 and DIARETDB1). MNIST and Fashion-MNIST are augmented by 5%, for the other two datasets we consider the no augmented version to be 50% (TUPAC16) and 90% (DIARETDB1) smaller.

The performances in Table 4.4 show that, CapsNet *without* data augmentation achieves a similar (TUPAC16, MNIST, Fashion-MNIST)

or even better (DIARETDB1) performance than ConvNets using data augmentation. All results are significant, the only Baseline for MNIST is comparable to the performance of CapsNet. These results confirm the benefits of equivariance over invariance.

4.3 Conclusion

In this work, we experimentally demonstrate the effectiveness of using CapsNet to improve CADx classification performance under medical data challenges. In particular, we demonstrate the increased generalization ability of CapsNets *vs.* ConvNets when dealing with the limited amount of data and class-imbalance. The performance improvement is a result of CapsNets equivariance modeling, that is, its ability to learn pose parameters along with filter weights. Together with the *routing-by-agreement* algorithm, this paradigm change requires to see fewer viewpoints of the object of interest, and therefore fewer images, in order to learn the discriminative features to classify them. We have also reported limitations to this otherwise general improvement of CapsNets over ConvNets, their improvement in performance is significant but has a limit that we observed for the more complex TUPAC dataset at 1% (5.5K training samples). Classification tasks where the global spatial structure plays a role can better exploit the advantages of CapsNets (DIARETDB1). One of the disadvantages of routing-by-agreement is that is slower than regular backpropagation, CapsNet with 8.2M parameters take about the same training time per epoch than Baseline with 35.4M (a ResNet-50 has 25.6M parameters). These architectures lack purposed layers, e.g. batch normalization, that could help to ease the convergence. Depending on the number of classes, CapsNet and Baseline need between 1-3 minutes per epoch, while LeNet runs in 1-2 seconds. Also, when visualizing the images reconstructed through the encoder-decoder branch (Fig. 4.3), we observe that they are blurry, especially for medical datasets with complex backgrounds. The fully-connected layers of this branch seem

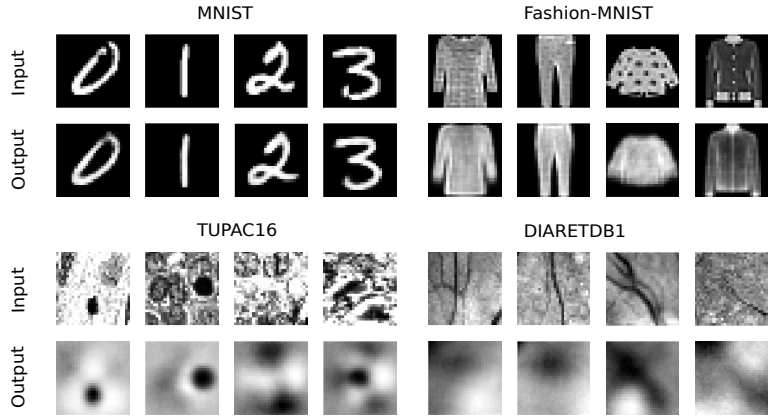

Figure 4.3. Test input images and their reconstructions.

to be good enough to regularize the parameter optimization but lose a lot of information. Our future work includes replacing these layers with deconvolutions to get a better insight into the learned latent space.

We recommend the use of capsule networks for medical datasets where the structure is important and patterns appear in different parts of the input images, as it is for retina. Our results confirm that they perform better than standard ConvNets for the limited amount of data, at least of the order of 10k. Another potential application would be the detection of rare diseases or segmentation due to the high performance under class-imbalance.

Part II

Analysis of Training Design

5 The Impact of Ordering and Pacing Training Samples

5.1 Introduction

Proximal femur fractures are a significant cause of morbidity and mortality, giving rise to a notable socioeconomic impact [RYY+15, GCG16]. Elderly population in the western world are especially affected. The incidence of femur fractures increases exponentially from an age of 65 and is almost doubled every five years.

Surgery is the most common and preferred treatment for proximal femur fractures [SSW+15]. The exact classification of the fracture is crucial for deciding the surgical procedure and choosing the surgical implant if needed. The Arbeitsgemeinschaft für Osteosynthesefragen (AO-Foundation) has established a hierarchical classification system for fractures of all bones based on radiographs. For proximal femur fractures, the AO classification has been beneficial, in terms of reproducibility, when compared against other systems such as the Jensen classification [BDE+15]. The AO standard follows a hierarchy according to the location and configuration of the fracture lines, see Fig. 5.1. Fractures of type-A are located in the trochanteric region,

and fractures of type-B are those affecting the area around the femur neck. Each type of fracture is further divided into 3 subclasses depending on the morphology and number of fragments of the fracture.

The ability to adequately classify fractures according to the AO standard based on radiographs is acquired through daily clinical routine in the trauma surgery department. Several years are needed until experienced trauma surgeons are significantly differentiated from residents. Inter-reader agreement varies from 66% among residents to 71% among experienced trauma surgeons [Zuc96]. To reach a precise classification, medical students and young trauma surgeons rely on a second opinion to choose the adequate treatment option for the patient. Our work aims to provide support as a computer-aided diagnosis (CAD) system capable of classifying radiographs.

Convolutional neural networks (CNNs) are nowadays the model of predilection for CAD. They have been rapidly integrated in numerous medical applications [SRG+16a, SWS17, GLS+18, BRPC+20, BLZ+18] due to their strong capacity to learn, directly from data, meaningful and hierarchical image representations. However, their feature extraction ability heavily depends not only on the optimization scheme but also on the training dataset. To be properly trained, CNNs need a large dataset representative of the population of interest [KSH12a].

In general, in medical image analysis tasks, acquiring reliable and clinically relevant annotated data remains a key challenge. Apart from the intra- or inter-expert disagreement, typically, manual annotations call for the time and effort of clinical experts. In addition, medical datasets usually suffer from class imbalance due to difficulties in collecting cases and the incidence of rare diseases. Finally, medical image data needs also dealing with proprietary and/or privacy concerns. As a result, these datasets generally exhibit three main challenging characteristics: (i) limited amounts of data, (ii) class-imbalance, and (iii) uncertain annotations.

The most common approaches to alleviate these challenges have been transfer learning [SRG+16a, ZLZ+17b, BRPC+20, SSY+19],

data augmentation [VV17a] and semi-supervised learning [SSCY19, SSY+19]. More recently, the attention has been shifted towards bootstrapping or weighting strategies [RCS+17], sample mining [XDS+19], active learning [SCN+18], and curriculum learning [TWH+18, YWL+19, MBN+18, JSMK+19].

The underlying intuition of strategies such as reordering, sampling or weighting, is that they can significantly impact the optimization of CNNs during training. Towards this objective, we reunite and formulate the above curriculum learning (CL) strategies to improve the performance of fine-grained proximal femur fractures classification, by dealing with the lack of large annotated datasets, class imbalance, and annotation uncertainty. Inspired by the concept of curriculum in human learning, CL presents the training samples to the algorithm in a meaningful order (often by difficulty from "easy" to "hard") and has been shown to avoid bad local minima and lead to an improved generalization [BLCW09].

Lately, training CNNs with ordered sequences has been shown to improve medical image segmentation by gradually increasing the context around the areas of interest [HGCB16, JGG+17, KDGBA19]. To the best of our knowledge, only few works have explored sample reordering for CAD with CNNS, for instance by extracting prior knowledge from radiology reports [TWH+18] or medical guidelines [JSMK+19].

The ordering can be either fixed (*e.g.* set heuristically by a "teacher" or domain-specific knowledge) or, in the absence of a-priori knowledge, a self-paced order [KPK10] derived from the algorithm's performance (*e.g.* the loss). Our unified CL formulation encompasses both approaches. We address the lack of prior knowledge to design an ad-hoc curriculum, by providing a ranking criterion based on uncertainty modelling. By using uncertainty to define our ranking, the classifier favors samples that it has not yet properly learnt, thus guiding it to explore "unseen" parts of the input space. We present three manners to actually implement the curriculum data sequencing. The first one is based on reordering the training set. The second uses a sampling strategy, *i.e.* selecting increasingly growing subsets. The

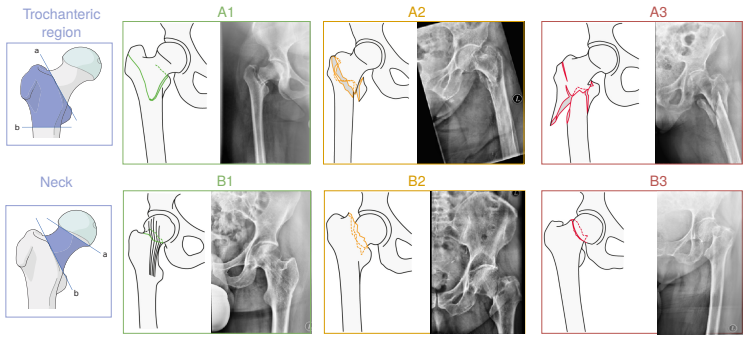

Figure 5.1. Examples of proximal femur fractures and their fine-grained AO classification, adapted from [KMA+18].

last one employs a weighting scheme to give different importance to the training samples.

To show the impact of our proposed method, we perform two types of experiments. First, on the challenging problem of multi-class classification of proximal femur fractures. This multi-class problem is inherently imbalanced, as the frequency of the classes reflects their incidence. Moreover, the adequate classification takes several years of daily clinical routine in the trauma surgery department, limiting the collection of annotations and leading to potentially noisy labels. Thus, to deepen the understanding of the method and to verify its effectiveness under these challenging data conditions, we design a series of experiments on the MNIST dataset, controlling the amount of data, class-imbalance, and label noise.

Contributions. In this work, we propose three CL strategies to automatically schedule the order and pace of the training samples for an improved multi-class classification. Our contributions are:

- We identify common curriculum learning elements among different data scheduling strategies, and present them within a unified formulation.

- We propose two types of novel ranking functions guiding the prioritization of the training data.
- We leverage domain-specific clinical knowledge to define the first scoring function.
- In absence of domain knowledge, we propose to estimate the ranking of the training samples by dynamically quantifying the uncertainty of the model predictions.
- We validate our strategies on a clinical dataset for the multi-class classification of proximal femur fractures.
- With a controlled experimental setting, we confirm that our method is useful in reducing the classification error under limited amounts of data, imbalance in the class distribution, and unreliable annotations. We give recommendations about the best approaches for each scenario.

This study is structured as follows. Section 6.2 covers CL related works that are relevant for the design of data schedulers. In Section 6.3, the details of our proposed formulation are presented. Section 6.4 describes the specifications of the experimental validation. Section 6.5 shows the classification performance. Section 6.6 discusses our findings, recommendations and future work. Finally, Section 6.7 summarizes our conclusions.

5.2 Related work

Recently, CL, self-paced learning (SPL), active learning (AL) and selection strategies have been studied to improve CNN-based image classification performance. These methods rely on ranking the training samples according to some criterion. In the following, we highlight some works that employ the two criteria related to our method: (i) domain-specific prior knowledge and (ii) data and model uncertainty.

Prior knowledge is leveraged in [TWH+18, JSMK+19, YWL+19, MBN+18] to design a curriculum for classification. Yang *et al.* [YWL+19] exploited SPL to handle class-imbalance, by combining the number of samples in each class and the difficulty of the samples, which is derived from the loss. Tang *et al.* [TWH+18] proposed to feed the images in order of difficulty based on severity-levels mined from radiology reports to improve the localization and classification of thoracic diseases. Jiménez-Sánchez *et al.* [JSMK+19] exploited the knowledge of the inconsistencies in the annotations of multiple experts and medical decision trees, to design a medical-based deep curriculum that boosted the classification of proximal femur fractures. Trying to mimic the training of radiologists, Maicas *et al.* [MBN+18] proposed to pretrain a CNN model with increasingly difficult tasks, before training for breast screening. The pretraining tasks were selected using teacher-student CL. In this work, we schedule our training data based on a scoring function that ranks the samples according to domain-specific prior knowledge or uncertainty. Different from previous works [TWH+18, JSMK+19, MBN+18], which only considered reordering the training set, here we investigate two further curriculum strategies, namely, subset sampling and weighting. Furthermore, solely Yang *et al.* [YWL+19] targeted one of the mentioned data challenges: class-imbalance, whereas we investigate as well noisy labels and limited amounts of training data.

The second criterion that we consider for defining a curriculum is uncertainty. The estimation of uncertainty provides a way of systematically defining the difficulty of the samples. Xue *et al.* [XDS+19] proposed online sample mining based on uncertainty to handle noisy labels in skin lesion classification. In their work, uncertainty is approximated through the classification loss. However, the most common methods for estimating classification uncertainty, in the context of deep learning, rely on Bayesian estimation theory, namely using Monte-Carlo (MC) dropout [GG16]. Uncertainty is probably the most frequent criterion in AL selection strategies. Recently, Wu *et al.* [WRL+18] combined uncertainty together with image noise into

their AL scheme to alleviate medical image annotation efforts. Uncertainty and label correlation are integrated in the sampling process to determine the most informative examples for annotation. AL pays attention to examples near the decision surface to infer their labels. Similarly, we aim to gradually move the classification decision border by adding examples of increasing ranking scores. We prioritize in our second scoring function the most representative samples, letting uncertainty guide their order, pace or weight. Although uncertainty has been used as sampling criterion for AL, we employ this information, for the first time, to rank and define our curriculum.

We validate our proposed curriculum strategies for the classification of proximal femur fractures. Whereas most of the previous work on femur fractures focuses on the binary fracture detection task [BZOR+19, CHL+19, WLC+19b], we target the more challenging multi-class classification according to the AO standard [KAS+17b, JSMK+19, JSKA+20].

Approaches to boost fracture classification accuracy comprise prior localization, transfer learning or medical knowledge. The localization of a region of interest before the classification of the full image has been studied either in a weakly-supervised [WLC+19b, KAS+17b] or in a supervised [JSKA+20] way. Knowledge transfer has been investigated across image domains, *i.e.* using ImageNet dataset for pretraining [UTG+19, BZOR+19], and across tasks, *i.e.* training first on body part detection (easier task) and then focusing on the hip fracture detection [CHL+19]. Medical knowledge has been proposed to train a hierarchical cascade of classifiers [TVM+20] or to schedule training data into a set of increasing difficulty [JSMK+19]. Tanzi *et al.* [TVM+20] relied on a cascade of classifiers. However, this kind of strategies are prone to propagate errors in multi-class classification. Furthermore, our CL approach does not rely on a complicated multistage scheme. We do not introduce any further complexity to the CNN.

In our previous work [JSMK+19], a series of heuristics, based on knowledge such as medical decision trees and inconsistencies in the

annotations of multiple experts, were proposed as a scoring function to boost fracture classification performance. Here, we further propose two more strategies, and also provide an alternative mechanism to rank the samples, based on prediction uncertainty, in case prior knowledge is unavailable.

5.3 Method

Given a multi-class image classification task, where an image x_i needs to be assigned to a discrete class label $y_i \in \{1, \ldots, T\}$, our training set is defined as $X = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. Assume a CNN model h with parameters θ is trained with stochastic gradient descent (SGD). During training, samples are typically randomly ordered. Our goal is to instead schedule the order and pace of the training data presented to the optimizer to better exploit the available data and annotations, and thereby improve the classification performance.

To learn the best CNN model h_{θ^*} from the input data, a common choice is to use empirical risk minimization:

$$\begin{aligned} \mathcal{L}(\theta) &= \tilde{\mathbb{E}}[L_\theta] = \frac{1}{N} \sum_{i=1}^{N} L_\theta(x_i, y_i) \\ \theta^* &= \arg\min_\theta \mathcal{L}(\theta) \end{aligned} \tag{5.1}$$

where $\tilde{\mathbb{E}}$ stands for the empirical expectation, L_θ is the loss function that measures the cost of predicting $h_\theta(x_i)$ when the correct label is y_i.

Optimization is conducted with SGD for a total of E epochs. Typically, the objective function L_θ is non-convex and is minimized in mini-batches of size B. Whereas convex learning is invariant to the order of sample presentation, CNNs are not. In the later case, the loss function usually presents a highly non-convex shape with many local minima, so the order of sample presentation affects learning, and thus, the final solution. It has been empirically shown that

the variance in the direction of the gradient step defined by easier examples is significantly smaller than that defined by difficult ones, especially at the beginning of training [NSW16, WCA18]. This suggests that favoring the easier examples may increase the likelihood to escape the attraction basin of an initial poor local minimum. Taking into account the mini-batches, we can rewrite Eq. (5.1) as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{j=1}^{N/B} \sum_{k=1}^{B} L_\theta(\hat{x}_{k,j}, \hat{y}_{k,j}), \tag{5.2}$$

where $\hat{x}_{k,j}$ is the k-th sample in the j-th batch, $\hat{x}_{k,j} = x_{k+(j-1)\cdot B}$, and $\hat{y}_{k,j}$ is the corresponding label.

We propose to modify Eq. (5.2) to schedule the training data. To do so, first, we formalize two types of scoring functions to assign a priority level to each data sample. The scoring is defined in Subsection 5.3.1 either according to domain-specific prior knowledge or to the samples' uncertainty measured with MC dropout. Then, in Subsection 5.3.2, we introduce the different components required for reordering, pacing, and weighting the training data. Finally, we cover the implementation details of the three variants of our unified CL formulation in Subsection 5.3.3.

5.3.1 Scoring function definition

The key element of our approach is the definition of the scoring function s or, equivalently, the curriculum probabilities p, which corresponds to normalized score function values. The formal definition of the curriculum probabilities is presented in Subsection 5.3.2. The curriculum allow us to sample the dataset and obtain a reordering function π that schedules the training samples. In this subsection, we present two alternative scoring functions. The first one is static and based on some initial (domain) knowledge, as in classical CL [BLCW09]. The second one is dynamic and based on the estimation of uncertainty, inspired by SPL [KPK10, WGY+19].

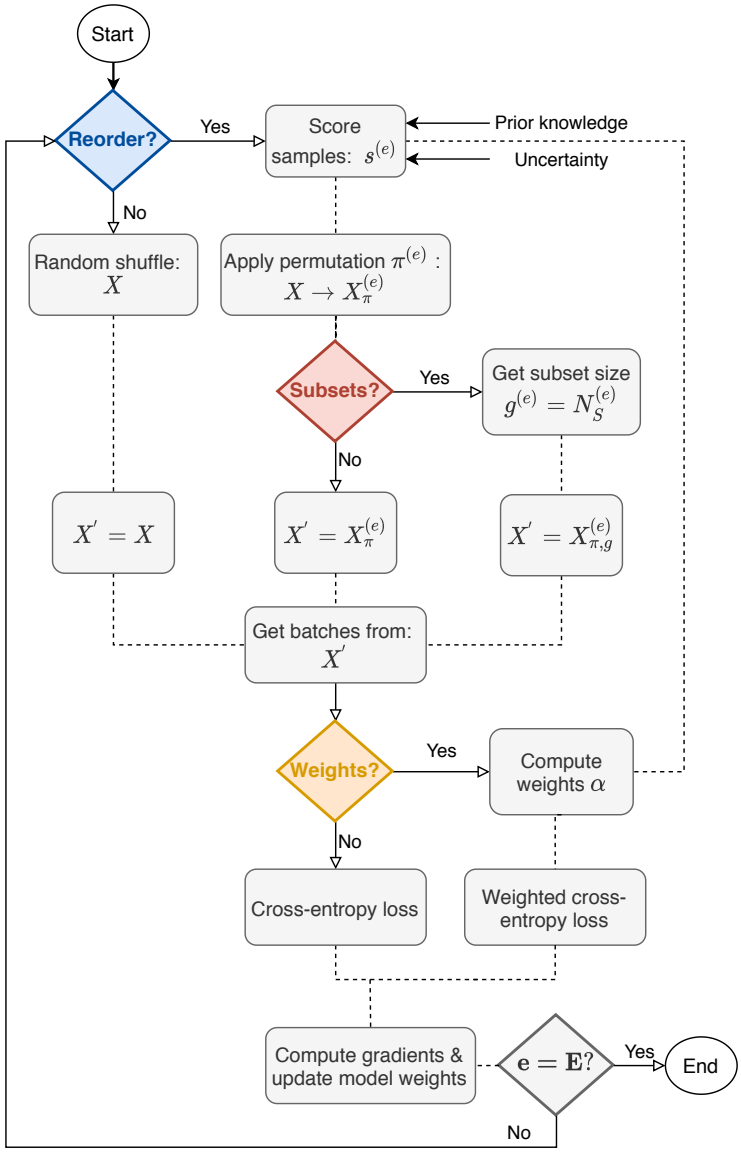

Figure 5.2. Diagram illustrating the components of the proposed unified CL method reuniting the three scheduling strategies: reorder, subsets, and weights. Straight lines are employed after a Yes/No junction because the flow is split. Otherwise, dotted lines are employed when there is no split.

Prior knowledge

In this scenario, the initial scoring $s^{(0)}$ and, thus, curriculum probabilities $p^{(0)}$, are specified based on domain prior knowledge. We assume in this variant that the scoring values are defined per class:

$$s^{(0)}(\cdot, y = t) = \omega_t, \tag{5.3}$$

where $t \in \{1, \ldots, T\}$ serves as index of the classes. ω_t is defined specifically for each task (or dataset). Once that the scoring values have been initialized, they can be kept fixed or decayed towards a uniform distribution [BLCW09]. In either case, as the curriculum probabilities are predetermined a priori in Eq. (5.3), we refer to this approach as static CL.

Prior knowledge can be obtained, for example, extracting keywords from medical reports [TWH+18], based on the frequency of samples [YWL+19, JSMK+19], employing medical classification standards or quantifying inconsistencies in the annotations [JSMK+19]. Specifically for this work, we define the initial probabilities for the proximal femur fracture images based on the Cohen's kappa coefficient [Hal12]. This statistic is used to measure the agreement of clinical experts on the classification between two readings. Basically, the kappa coefficient quantifies the ratio between the observed and chance agreement. To better understand and illustrate the potential of CL, we also analyze our method on MNIST dataset. In this case, we extract prior knowledge by ranking the per-class F_1-score performance after few epochs of training. The exact values used for our experiments are specified in Subsection 6.4.3.

Uncertainty estimation

In absence of domain knowledge, we propose to estimate the priority of the training samples by dynamically quantifying the uncertainty of the model predictions. Uncertainty provides a way of systematically ranking the training samples based on the model's agreement on the predictions, with the benefit of not requiring any prior knowledge. At

each epoch e, we compute the uncertainty in predicting a sample x_k, and use such uncertainty as its scoring value s_k. See Subsection 5.3.2 for the definitions of x_k and s_k. The goal is to emphasize samples with high information gain at early stages of training, *i.e.* to rapidly reduce the error in highly-misleading samples.

To estimate the uncertainty of the model predictions, we employ MC dropout [GG16]. In this training regime, each epoch includes two stages [LLT19]: uncertainty estimation and label prediction. In the uncertainty estimation stage, we perform L stochastic forward passes on the model under random dropout. The L estimators are used to measure the uncertainty of the output of the model. In the prediction stage, a single forward pass is performed. Then, the classification loss is used to measure the difference between the prediction and the label.

Let $\sigma \in \mathbb{R}^T$ be the (softmax) output of the CNN. This output represents the probability distribution of the predicted label over the set of the possible classes for sample x, *i.e.* $P(y = t \mid x, \theta) := \sigma_t$. We measure uncertainty as the entropy [Sha48] of the output distribution, *i.e.* predictive entropy:

$$H(y|x, \theta) = -\sum_{t=1}^{T} P(y = t \mid x, \theta) \cdot \log P(y = t \mid x, \theta). \tag{5.4}$$

This measurement helps to discriminate points that are far from all training data, yet the model assigns high confident prediction (low predictive entropy). We aim to minimize the effect of these samples, with a small weight or bringing them at a later stage in training [SG18].

The output distribution $P(y = t|x, \theta)$ can be approximated using MC integration:

$$\tilde{P}(y = t \mid x, \theta) = \frac{1}{L} \sum_{l=1}^{L} P(y = t \mid x, \theta_l), \tag{5.5}$$

where $P(y = t \mid x, \theta_l)$ is the probability of input x to take class t with model parameters $\theta_l \sim q(\theta)$, with $q(\theta)$ being the (dropout)

variational distribution. We set the scoring function to be the estimated predictive entropy, computed from the MC estimated output distribution $\tilde{\sigma}_t = \tilde{P}(y = t \mid x, \theta)$:

$$s = -\sum_{t=1}^{T} \tilde{\sigma}_t \cdot \log \tilde{\sigma}_t. \tag{5.6}$$

By assigning low scoring values to predictions with low predictive entropy, we decrease the priority of samples with low information gain. Note that in contrast with Eq. (5.3), here, the scoring elements s_i are defined independently for each sample, and updated after each epoch. Only few works measure uncertainty while learning the classification task [GGG+19]. To the best of our knowledge, our proposed dynamic uncertainty-driven curriculum strategy is novel for CAD.

5.3.2 Data scheduler

In the following, we define the scheduling elements required for reordering and pacing our training data: a scoring function s, curriculum probabilities p, a permutation function π, a pacing function g, and a weighting function α. The data scheduler takes as input the training set X, the scoring and pacing functions, s and g, respectively, and it outputs the reordered set/subset, partitioned in mini-batches. All components are updated at each epoch e.

- The *scoring function* $s : \mathcal{X} \to \mathbb{R}$ ranks the curriculum priority of each training pair. The curriculum priority can take various forms, such as difficulty or prediction disagreement. An example (x_i, y_i) has higher priority than example (x_j, y_j) if $s(x_i, y_i) > s(x_j, y_j)$. We define $s_i = s(x_i, y_i)$ and, in an abuse of notation, use s to denote both the scoring function and the vector $(s_1, \ldots, s_N)$.

- The *curriculum probabilities* p are obtained by normalizing the score function values (while preserving the order and ensuring

they add up to 1). For example, one can choose $p_i = s_i/||s||_1$, assuming $s_i \geq 0$. A pair (x_i, y_i) is more likely to be presented earlier to the optimizer than a pair (x_j, y_j) if $p_i > p_j$.

- The *reordering function* $\pi : [1, \ldots, N] \rightarrow [1, \ldots, N]$ is a permutation. It is determined by resampling without replacement X according to the curriculum probabilities p.

- The *pacing function* $g : \mathbb{N} \rightarrow \mathbb{N}$ controls the learning speed by presenting growing subsets of data. The batch size B is kept fixed. The non-decreasing mapping g determines the subset size $N_S \leq N$ at each training epoch e, *i.e.* $g(e) = N_S^{(e)}$.

- The *weighting function* $\alpha : \mathcal{X} \rightarrow \mathbb{R}$ favors the samples that have higher priority according to the curriculum probabilities. These per-sample weights are applied directly to the classification loss.

Taking into account the scheduling elements introduced, we can rewrite the optimization loss at epoch e as:

$$\mathcal{L}_\theta^{(e)} = \frac{1}{N_S^{(e)}} \sum_{j=1}^{N_S^{(e)}/B} \sum_{k=1}^{B} \hat{\alpha}_{k,j}^{(e)} \cdot L_\theta(\hat{x}_{k,j}^{(e)}, \hat{y}_{k,j}^{(e)}), \tag{5.7}$$

where $\hat{x}_{k,j}^{(e)} = x_{\pi^{(e)}(k+(j-1)\cdot B)}$ corresponds to the k-th sample from the j-th batch at epoch e after reordering π. The same relation follows for its corresponding label and weight, $\hat{y}_{k,j}^{(e)}$ and $\hat{\alpha}_{k,j}^{(e)}$, respectively. We will drop superscript (e) when no confusion arises. Also, we simplify notation and use x_k (and y_k, α_k) to refer to a given (already reordered) sample (and label, weight). This equation encompasses the three main curriculum strategies from the literature: reordering, increasing subsets, and weighting.

5.3.3 Scheduling data with curriculum learning

In practice, any curriculum is implemented by assigning a predefined or estimated probability p_i to each training pair (x_i, y_i), as described

in Subsection 5.3.1. Fig. 5.2 visualizes the data flow in the different scheduling strategies, each of them being depicted by a diamond shape: reorder, subsets, and weights. The scoring function s and curriculum probabilities p are common to the three scheduling approaches, whereas the reordering function π is used in the reorder and subset strategies.

The first mechanism, *reorder*, presents the samples to the optimizer in a "smart" probabilistic order, instead of the typical random permutation. This strategy aims to deal with low-priority cases at a later stage of training [BLCW09, HW19, JSMK+19]. At the beginning of every epoch e, the training set $X = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ is permuted to $X_\pi^{(e)} = \{(x_{\pi^{(e)}(1)}, y_{\pi^{(e)}(1)}), \ldots, (x_{\pi^{(e)}(N)}, y_{\pi^{(e)}(N)})\}$ using the reordering function $\pi^{(e)}$. This mapping results from sampling the training set according to the curriculum probabilities $p^{(e)}$ at the current epoch e. Mini-batches are formed from $X_\pi^{(e)}$.

The second method, *subsets*, builds upon the reordered training set and selects gradually increasing subsets at every epoch. The purpose is to reduce the effect of outliers at the beginning of training [HW19, XDS+19, WGY+19]. Mini-batches are obtained from $X_{\pi,g}^{(e)} \subseteq X$, where $X_{\pi,g}^{(e)}$ are the first $N_S^{(e)}$ pairs of $X_\pi^{(e)}$. The subset size at every epoch $N_S^{(e)}$ is determined by the pacing function g. For simplicity, in our experiments we choose g to be a staircase function:

$$g(e) = N_S^{(e)} = \begin{cases} N_S^{(0)} + e \cdot \Delta & if \quad 1 \leq e < E_S \\ \\ N & if \quad e \geq E_S \end{cases} \tag{5.8}$$

where $\Delta = (N - N_S^{(0)})/E_S$, $N_S^{(0)}$ is a predefined initial subset size, and E_S is the number of epochs before considering the whole training set.

A counter τ_i is introduced to track the selected pairs. Their scoring vector is decreased, thus favoring new pairs in the subsequent epoch. We choose to update the scoring vector using an exponential decay:

$$s_i^{(e)} = s_i^{(e-1)} \cdot \exp(-\tau_i^2/10) \quad e = 1, \ldots, E. \tag{5.9}$$

The third approach, *weights*, assigns scalar weights to training samples based on their curriculum probabilities [WGY+19]. We propose to weight the classification loss L_θ of each training sample in Eq. (5.7), in the form of a weighted cross-entropy loss. The role of the weights is to decrease the contribution to the classification loss of samples with low priority. We choose the weights $\hat{\alpha}_{k,j}$ to correspond to a per-batch normalization of the curriculum probabilities:

$$\hat{\alpha}_{k,j}^{(e)} = \frac{p_{k+(j-1)\cdot B}^{(e)}}{\max_m p_{m+(j-1)\cdot B}^{(e)}} = \frac{\hat{p}_{k,j}^{(e)}}{\max_m \hat{p}_{m,j}^{(e)}}. \tag{5.10}$$

When the curriculum is driven by uncertainty, the resulting approach is similar to boosting [FSA99]. In the boosting method, misclassified examples are given a higher weight than correctly classified ones. This is known as "re-weighting". Following the same principle, we use the uncertainty at every epoch, in our curriculum data scheduler, to update the values of the weights.

5.4 Experimental validation

In order to validate the positive effect of data scheduling on the classification performance, we perform experiments on two types of image databases: (i) a real in-house dataset of a moderate size and naturally suffering from imbalance and noisy labels, and (ii) the MNIST dataset. The second one is used for additional analysis under controlled experiments to further illustrate the potential of CL.

5.4.1 Datasets

Proximal femur fractures. Our clinical dataset consists of anonymized X-rays of the hip and pelvis collected at the trauma surgery department of the Rechts der Isar Hospital in Munich. Images of 2500×2048 pixels were gathered from a group of 780 patients. Each patient study

contained one or two radiographs. Most of the images were anterior-posterior (a-p), only 4% were side view. The collection of these radiographs was approved by the ethical committee of the Faculty of Medicine from the Technical University of Munich, under the number 409/15 S. The dataset consists of 327 type-A, 453 type-B fractures and 567 non-fracture cases. Class labels were assigned by clinical experts according to the AO classification standard [KMA+18]. Each type of fracture is further divided into 3 subclasses depending on the morphology and number of fragments of the fracture, see Fig. 5.1. Subtypes of the fracture classes are highly unbalanced, reflecting the incidence of the different fracture types. In particular, the number of images for the subclasses is as follows: type-A (114, 197, 16), and type-B (79, 241, 133). Clinicians also provided square bounding box annotations containing the head and neck of the femur. We leveraged these annotations, cropped and resized the image to 224×224 pixels. The dataset was split patient-wise into three parts with the ratio 70%:10%:20% to build respectively the training, validation and test sets. We evaluate the classification performance of the 3-class (type-A or type-B and non-fracture) and 7-class (fracture subtypes and non-fracture) classification tasks. The train, validation and test distributions were balanced between fracture type-A, type-B, and non-fracture cases. To achieve an equal proportion of subtype representation (of approximately 12%), data augmentation techniques were used. Specifically, techniques such as translation, scaling and rotation were combined.

MNIST. The MNIST handwritten digit database is publicly available[1]. It has a training set of 50000 examples and a validation and test sets of 10000 examples each. Classes are equally represented.

[1] http://yann.lecun.com/exdb/mnist/

5.4.2 Experimental Setting

We perform a comparative evaluation of the classification task with five series of experiments. Our method is contrasted against its "anti-" approach, *i.e.* the curriculum probabilities are complemented, "random" criterion, *i.e.* the curriculum probabilities are assigned randomly, and the "baseline" model. The baseline model does not consider any data scheduling elements, and it is trained on randomly shuffled versions of the whole training set.

In the first series of experiments, we examine the performance of our method driven by prior knowledge. In the second series, we consider the use of uncertainty to overcome the lack of prior knowledge. Our clinical dataset inherently suffers from class-imbalance, unreliable annotations and a limited size. The 7-class discrimination task is challenging as reflected by i) the existing intra- and inter-expert agreement (66% among residents *vs.* 71% among experienced trauma surgeons); and ii) the long and shallow learning curve of young trauma surgery residents who acquire the classification skills during the daily routine. For the remaining experiments, we employ MNIST, as a controlled environment, to investigate such challenging scenarios. In the third series, we evaluate the classification performance when training with limited amounts of data. In the fourth series, we present the results that deal with class-imbalance. Finally, in our last series of experiments, we discuss and show the performance under the presence of label noise.

5.4.3 Implementation details

Architectures and optimization hyperparameters. We train our models 10 times for 30 epochs, with an early stopping criterion of no improvement in the validation set for 20 epochs. For the digit recognition task, we use an upgraded ConvPool-CNN-C [SDBR14] proposed by [LA17], illustrated in Fig. 5.3 of Suppl. Material. This architecture replaces pooling layers by convolutional layers with a stride of two. Besides, the small convolutional kernels greatly re-

duce the number of parameters of the network. It yielded competitive performance on several object recognition datasets (CIFAR-10, CIFAR-100, ImageNet). For the fracture classification, we deploy a ResNet-50 [HZRS16b] pretrained on the ImageNet dataset, on account of the limited size of our dataset and the benefits of transfer learning [SRG+16a, SSY+19]. We limit our evaluation to those two CNNs, since Weinshall *et al.* [WCA18] reported that CL lead to an improved generalization performance with both 'small' and 'large' architectures. For both architectures, we use a mini-batch size of 64, an initial learning rate of 1e−3, and a dropout rate for the fully connected layer of 0.9 (0.7 for uncertainty estimation). Our ResNet-50 is trained with SGD and a momentum of 0.9. The learning rate is decayed by a factor of 10 every 10 epochs. ConvPool-CNN-C is trained with Adam. For the weighting strategy, since the batch size is directly related to the computation of the sample weights, we evaluated different batch sizes (16, 32, and 64). We found that the curriculum is robust, achieving the lowest standard deviation for $B = 64$ (see Table 5.8 in Suppl. Material). For the subsets strategy, we choose as hyperparameters: the warm-up epochs $E_S = 10$ and the initial subset size $N_S^{(0)}$ to 25% of the training data size at each scenario. We evaluated several warm-up epochs $E_S = \{5, 10, 20\}$ and sizes for the initial subset $N_S^{(0)} = \{25\%, 40\%\}$. Results for the different configurations were comparable (see Tables 5.6-5.7 in Suppl. Material).

Prior knowledge.

- Proximal femur fractures. In this setting, we leverage, as prior knowledge the intra-reader agreement from a committee of experts: a trauma surgery attendant with one year experience, a trauma surgery attending and a senior radiologist. The scoring values for the seven classes are the following:

$$\omega = (0.69, 0.56, 0.62, 0.60, 0.56, 0.38, 0.92). \tag{5.11}$$

  These values correspond to the multi-read kappa agreement described in Results section [JKA+19].

- MNIST. In absence of domain-specific knowledge, a CNN is trained for 5 epochs. After observing the F_1-score of each of the classes, weights are assigned, by ranking the classes from easiest (highest F_1-score) to hardest (lowest F_1-score). Then, training is restarted from scratch using these particular weights. We specify the values for the experiments with limited amounts of data $\omega_{limited}$, under class-imbalance $\omega_{imbalance}$, and with noisy labels ω_{noise}:

$$\omega_{limited} = (7, 10, 5, 4, 9, 1, 8, 6, 2, 3) \tag{5.12}$$

$$\omega_{imbalance} = (3, 10, 7, 8, 5, 6, 9, 4, 1, 2) \tag{5.13}$$

$$\omega_{noise} = (8, 10, 9, 7, 5, 1, 2, 3, 4, 6). \tag{5.14}$$

Prior knowledge		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	56.62	34.56	**<u>68.93*</u>**	58.90	50.89	**66.50***	58.26	55.20	**64.65***
3-class	81.71	60.46	**<u>86.23*</u>**	80.82	75.64	**84.69***	80.66	75.33	**85.66***
Uncertainty		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	56.62	61.29	**64.70***	58.90	62.06	**<u>65.51*</u>**	58.26	58.29	**62.29***
3-class	81.71	82.48	**84.38***	80.82	82.79	**<u>84.90*</u>**	80.66	82.69	**82.96***

Table 5.1. Fracture classification results over 10 runs: mean F_1-score. The highlighted indices in bold correspond to the best metric per curriculum method. The underlined values correspond to the best metric per scenario, *i.e.* 3-class (type-A or type-B and non-fracture) and 7-class (fracture subtypes and non-fracture) classification. Statistical significance with respect to baseline is marked with *.

5.5 Results

5.5.1 Prior knowledge-driven CL

We evaluated the performance of the classifier with our data scheduler and verified that establishing a curriculum based on prior knowledge

Prior knowledge		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.19	9.28	**5.46***	5.60	13.17	**4.29***	8.01	5.78	**5.35***
50% MNIST	3.36	5.21	**2.53***	3.96	4.21	**2.05***	4.10	4.11	**2.96***
100% MNIST	1.67	2.53	**1.32***	1.96	1.78	**1.17***	1.98	1.79	**1.32***
Uncertainty		**Reorder**		**Subsets**			**Weights**		
	Baseline	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.19	8.94	**4.42***	5.60	8.85	**3.69***	8.01	8.50	**5.62***
50% MNIST	3.36	3.23	**3.04**	3.96	4.21	**2.15***	4.10	4.77	**3.21**
100% MNIST	1.67	2.29	**1.45**	1.81	2.02	**1.17***	1.99	1.66	**1.33***

Table 5.2. Digit classification results over 10 runs: mean error rate (%). The highlighted values in bold correspond to the best metric per curriculum method. The underlined values correspond to the best metric per scenario, *i.e.* percentage of data. Statistical significance with respect to baseline is marked with *.

		Reorder		**Subsets**		**Weights**	
	Baseline	Prior K.	Uncertainty	Prior K.	Uncertainty	Prior K.	Uncertainty
Class-imbalance	2.53	2.08	2.05	**1.79**	2.08	2.31	2.22
Label Noise	9.46	8.76	8.42	8.28	7.24	8.49	**5.42**

Table 5.3. Comparison of curriculum strategies driven by prior knowledge and uncertainty, under class-imbalance and label noise for the MNIST dataset. Mean error rate (%). The highlighted values in bold correspond to the best strategy per scenario.

is a good and suitable option to improve classification performance. Results for proximal femur fracture classification are summarized in Table 5.1-top, and for digit recognition in Table 5.2-top. We found that the three variants helped to improve the performance of the two datasets. In contrast with the anti-CL approach, accuracy was in every case increased with respect to the baseline.

For MNIST, we found that training starting with an easy subset, and gradually increasing the subset by adding more difficult samples was the best strategy for the three scenarios as shown in Fig. **??**-a. A comparable improvement with respect to the baseline was found

when we introduced the decay of Eq. (5.9) in reorder strategy and performed instead sampling with replacement.

For fracture classification, the F_1-score for 7-class was improved up to 15% compared to the baseline. This score is comparable to state-of-the-art results [JSMK+19] and experienced trauma surgeons [vERMR10]. Tanzi *et al.* [TVM+20] reported an average F_1-score of 0.76 for the easier 5-class classification task on a private dataset. The authors did not consider the subcategories of type-B fracture. We hypothesize that by reordering the whole training set instead of using subsets, we improve diversity by including the more challenging fine-grained fractures classification task. Furthermore, as specified in Subsection 6.4.3, the CNN for fracture classification was pretrained, whereas for digit recognition the CNN was trained from scratch. From the results in Table 5.1, we can say that our method is compatible with transfer learning.

5.5.2 Uncertainty-driven CL

Here, assuming lack of prior knowledge, we confirmed that uncertainty estimation can guide the data scheduling. Results are presented in Table 5.1-bottom and Table 5.2-bottom for fractures and MNIST, respectively. For the fine-grained 7-class proximal femur fractures classification, the F_1-score was improved up to 16% compared to the baseline. In this case, we found that weighting the samples was not as beneficial as reordering or sampling subsets. For digit classification, the error rate was reduced up to 30%, see Fig. **??**-b. Anti-CL leading to a better performance than the baseline is a behaviour also reported in [BLCW09]. Furthermore, we found that this behaviour was sporadic and not statistically significant with respect to the baseline, whereas the CL approach was consistent and statistically significant.

5.5.3 Limited amounts of data

Table 5.2 shows the digit recognition performance when restricting the amount of training data to 30% and 50%. When employing our curriculum strategies, the error rate for digit classification is reduced in all cases. We found that employing subsets in the first epochs based on uncertainty was the best strategy. Moreover, the effect of our curriculum approach was more evident on the more challenging scenario. The error rate was reduced by up to 59% training with only only 30% of the data The fact that our CL schemes are beneficial with limited amounts of annotated data makes it appealing for annotation-efficient learning on other medical image datasets. Interestingly, we found that when training with only 30% of data, the use of random subsets also reduced the error rate. This behaviour goes along with some findings about training with partial data [MA18].

5.5.4 Class-imbalance

We evaluated our proposed curriculum method in a controlled experiment under class-imbalance with the MNIST dataset. Specifically, the number of examples of two classes (digits 1 and 7) are limited to 30% of the available cases. Results in Table 5.3 show that our approach can cope with class-imbalance and improved over the baseline result. Similar to the experiment with limited amounts of data, the use of high-priority subsets, selected based on prior knowledge or uncertainty, was the best approach. The subsets approach reduced the error rate from 2.53% to 1.79%.

5.5.5 Noisy labels

Using MNIST and a controlled setting, we corrupted a randomly selected 30% of training labels by assigning to them the subsequent label digit, *i.e.* zeroes become ones, ones become twos, *etc.* Table 5.3 reports the mean error rate (%) when evaluating the digit classification. We found that our three CL schemes were effective to deal

with noisy labels and beat the baseline. The benefit of curriculum under noisy regime is also confirmed on a recent work by Wu *et al.* [WDN21]. We investigated the role of the curriculum probabilities in the weights strategy. We found that the uncertainty-driven curriculum assigns high-value weights to a larger amount of clean samples than the 'random' strategy (see Fig. 5.4 of the Supplementary Material). Therefore, globally, the weighting curriculum gives more importance to cleaner samples. In this case, prior knowledge was not as beneficial as the estimation of model prediction uncertainty. The best variant was using uncertainty to weight the classification loss, reducing the error rate by 43%. The fact that uncertainty performed better than prior knowledge was expected, since noise may affect individual samples and not entire classes. It is more reasonable to use a scoring function that independently affects the samples. Moreover, weighting seemed to be the stronger strategy to remove or reduce the influence of the flawed labels.

5.6 Discussion

In this work, we bring together several ideas from the literature and present them into a unified CL formulation. We experimentally demonstrate the effectiveness of ranking and scheduling training data for the challenging multi-class classification of proximal femur fractures. Most of the previous work [BZOR$^+$19, CHL$^+$19, WLC$^+$19b] only target the fracture detection task, and Tanzi *et al.* [TVM$^+$20] does not obtain the same level of granularity. Our CL schemes achieve state-of-the-art results on the 7-class classification task. Furthermore, we also show the benefits of our CL strategies in a controlled set-up with MNIST dataset, specifically, under demanding scenarios such as class-imbalance, limited amounts of data and noisy annotations.

Inspired by classical CL, we leveraged prior knowledge to define the data scheduling elements. In our formulation, this prior knowledge only requires defining a scalar value per class. In case of mul-

tiple experts annotating the dataset, this knowledge can be derived from their intra- or inter-expert variability, or by asking the experts about the perceived difficulty of each class. One limitation of this approach is that the use of prior knowledge at the class level may be less informative for the CNN than at sample level. When prior knowledge is not available, we have shown that uncertainty can be used to guide the optimization. We used MC dropout to estimate uncertainty. This has the advantage of not requiring any change in the CNN architecture, but it is computationally demanding. Indeed, the training time is doubled in this variant. Instead, one could investigate the use of a Dirichtlet distribution to parametrize the output of the network. Then, the behavior of such predictor could be interpreted from an evidential reasoning perspective, such as in subjective logic [SKK18, Jøs18]. Future research directions for defining the scoring function could be based on other uncertainty measures such as quantifying out-of-distribution samples [HG16] or evidence theory [SKK18]. We restricted our study to the predictive entropy of the model, which includes both aleatoric and epistemic uncertainty. We reckon that assessing separately each type of uncertainty could be advantageous for some applications. Moreover, if training time is not a concern, uncertainty does not only rank at the class but at the sample level. This scoring function is more appropriate for noisy annotations, since noise may affect individual samples and not entire classes. It could also be interesting to investigate the CNN behaviour when using prior knowledge alternatives at sample level, rather than at class level.

We evaluated three CL variants that consisted of reordering the whole training set, sampling subsets of data, or individually weighting training samples. Our CL schemes are compatible with any architecture and SGD training [WCA18]. They only require domain-specific knowledge or the estimated uncertainty for the definition of the scoring function, hence the curriculum. The reordering and subsets performances are very similar but if the dataset is too complex for the amount of available data (fractures), it seems better to keep the en-

tire training set. We found similar performance when the curriculum probabilities were decayed towards a uniform distribution [BLCW09] or maintained stable in our reorder and weights variants. Regarding the latter, we have proposed a simple and effective weighting scheme. In future work, we plan to explore other weighting strategies, *e.g.* the focal loss [LGG+17], which is well suited for class-imbalance scenarios, and the large margin loss [EKM+18], which has been shown beneficial under limited amounts of data and when noisy labels are present.

5.7 Conclusions

In this work, we have designed three CL strategies for the multi-class classification of proximal femur fractures. We validated the benefits of our approach reaching a performance comparable to state-of-the-art and experienced trauma surgeons. We have identified common scheduling elements in the literature and unified their formulation in our approach. We have proposed two types of ranking functions to prioritize training data, leveraging: prior knowledge and uncertainty. In controlled experiments with the MNIST dataset, we have shown that the proposed method is effective for datasets with class-imbalance, limited or noisy annotations. From our experiments, we can conclude that for datasets of limited size or under the presence of class-imbalance, the use of the subsets variant can lead to an improved classification performance. One can either exploit prior knowledge to achieve a better performance, or if the computational cost is not an issue, leverage uncertainty. In the case of unreliable labels, we found that the more advantageous approach is the combination of weights with uncertainty.

Supplementary Material

Prior knowledge	Reorder		Subsets			Weights		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	**5.16E-04**	**3.03E-09**	1.33E-01	5.61E-02	**1.47E-06**	3.26E-01	6.34E-01	**1.48E-06**
3-class	**8.78E-04**	**5.43E-05**	6.02E-01	7.20E-02	**1.38E-02**	4.73E-01	**1.07E-02**	**1.57E-04**
Uncertainty	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
7-class	**7.10E-03**	**7.63E-05**	1.33E-01	**3.49E-03**	**2.27E-05**	3.26E-01	4.11E-01	**4.94E-05**
3-class	4.80E-01	**1.54E-02**	6.02E-01	4.37E-01	**2.98E-03**	4.73E-01	3.10E-01	**1.97E-02**

Table 5.4. Statistical significance analysis for proximal femur fracture experiments. T-test with respect to baseline. P-values below 0.05 are bold-faced.

Prior knowledge	Reorder		Subsets			Weights		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	9.40E-01	**8.70E-04**	**1.55E-04**	1.62E-01	**3.52E-06**	3.29E-01	**1.08E-04**	**6.03E-05**
50%MNIST	**4.58E-04**	**8.37E-04**	7.19E-02	8.59E-02	**6.83E-05**	2.04E-01	5.69E-02	**7.23E-03**
100%MNIST	**3.75E-03**	**1.22E-02**	3.27E-01	3.88E-01	**5.98E-04**	**3.09E-02**	3.83E-01	**9.95E-03**
Uncertainty	**Reorder**		**Subsets**			**Weights**		
	Anti-CL	CL	Random	Anti-CL	CL	Random	Anti-CL	CL
30% MNIST	8.38E-01	**7.16E-06**	1.55E-04	7.77E-01	**5.79E-07**	3.29E-01	5.01E-01	**3.11E-05**
50% MNIST	5.87E-01	3.77E-01	7.19E-02	1.15E-01	**1.86E-04**	2.04E-01	8.47E-02	6.43E-02
100% MNIST	**1.25E-02**	6.41E-02	3.27E-01	**3.44E-02**	**3.56E-03**	**3.09E-02**	7.11E-01	**2.49E-02**

Table 5.5. Statistical significance analysis for MNIST experiments. T-test with respect to baseline. P-values are reported.

Prior knowledge-driven CL		**Uncertainty-driven CL**	
$N_S^{(0)} = 25\%$	$N_S^{(0)} = 40\%$	$N_S^{(0)} = 25\%$	$N_S^{(0)} = 40\%$
66.50 (66.02) ± 2.00	65.39 (65.76) ± 2.23	65.51 (66.32) ± 3.37	64.99 (65.63) ± 2.30

Table 5.6. F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the subsets strategy with different initial subset sizes $N_S^{(0)}$.

Prior knowledge-driven CL			**Uncertainty-driven CL**		
$E_S = 5$	$E_S = 10$	$E_S = 20$	$E_S = 5$	$E_S = 10$	$E_S = 20$
63.68 (63.42) ± 3.15	66.50 (66.02) ± 2.00	66.09 (64.04) ± 1.24	65.30 (65.78) ± 3.12	65.51 (66.32) ± 3.37	66.42 (66.68) ± 1.95

Table 5.7. F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the subsets strategy with different number of epochs E_S before considering the whole training set.

Prior knowledge-driven CL			Uncertainty-driven CL		
$B = 16$	$B = 32$	$B = 64$	$B = 16$	$B = 32$	$B = 64$
65.35 (65.02) ± 2.97	65.69 (65.95) ± 2.11	64.65 (64.04) ± 1.56	64.66 (65.76) ± 2.27	66.92 (66.69) ± 2.18	62.60 (62.96) ± 1.63

Table 5.8. F_1-score for the 7-class fracture classification, mean (median) and standard deviation for the weights strategy with different batch sizes.

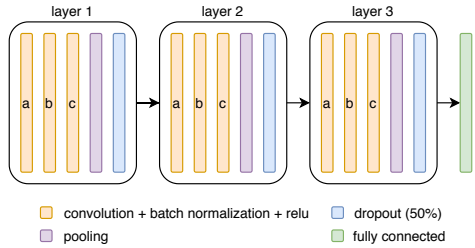

Figure 5.3. Network architecture employed for the experiments with the MNIST dataset.

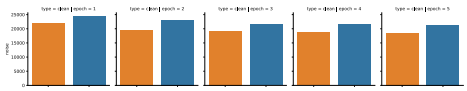

Figure 5.4. Analysis of weights strategy under label corruption for MNIST dataset. Number of samples with a weight higher than the mean weight at that epoch. Random criterion and uncertainty are depicted in orange and blue, respectively.

6 Memory-aware Curriculum Federated Learning

6.1 Introduction

Breast cancer is the most commonly occurring type of cancer worldwide for women [SFS+21]. Early detection and diagnosis of breast cancer is essential to decrease its associated mortality rate. The medical community recommends regular screening with X-ray mammography imaging for its early detection and follow-up. High-resolution images showing tissue details need to be analyzed to spot abnormalities and to provide a precise diagnosis. Despite high incidence (*i.e.* 12%)[SFS+21], the extensive breast cancer screening results predominantly in negative samples. Such class imbalance can be problematic for learning-based Computer-Aided Diagnosis (CAD) systems. A potential solution to mitigate the existing class-imbalance and to increase the size of the annotated dataset is to employ data coming from multiple institutions. However, sharing medical information across (international) institutions is challenging in terms of privacy, technical and legal issues. Secure and privacy-preserving machine learning offers an opportunity to bring closer patient data protection and data usage for research and clinical routine purposes.

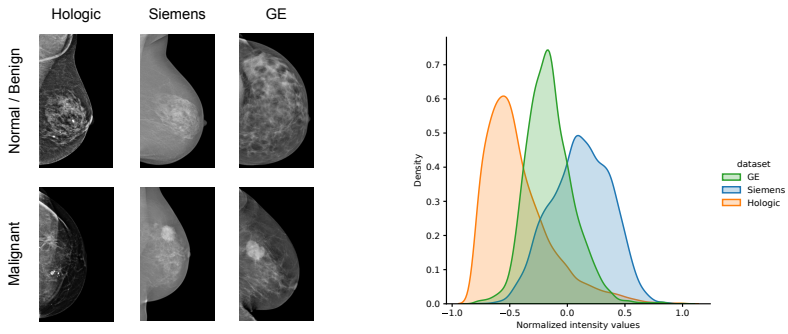

Figure 6.1. Left: Exemplary mammograms of benign and malignant cases. Right: pixel-intensity distributions of different sites.

Federated Learning (FL) aims to train a machine learning algorithm across multiple decentralized nodes holding locally the data samples, *i.e.* without exchanging them. Training such a decentralized model in a FL setup presents three main challenges: (i) system and statistical heterogeneity, (ii) data protection, and (iii) distributed optimization. We deal with the three challenges for breast cancer classification in the context of FL.

The first challenge concerns system and data heterogeneity. For the same imaging modality, different system vendors produce images following significantly different intensity profiles. To cope with such diversity, recent works [PHZS19, LJZ+21] have proposed to integrate Unsupervised Domain Adaptation (UDA) into the FL framework. UDA methods force the model to learn domain-agnostic features through adversarial learning [PHZS19] or a specific type of batch normalization [LJZ+21]. In this work, we follow an UDA adversarial approach to handle non-IID data.

To address the second challenge, data protection, cryptographic techniques [BIK+17] or differential privacy [DKM+06, DR+14] are employed. Differential privacy perturbs each local model parameters by purposely adding noise before uploading them to the server for aggregation. We leverage differential privacy for data protection in

our method.

The third challenge concerns the distributed optimization in the FL setting. Individual models are trained locally on private data and the central server is responsible for the global aggregation of the local updates. Usually, the communication of the local models to the server occurs a certain number of times every epoch. Therefore, we propose a novel curriculum learning approach that provides a meaningful order to the samples.

Contributions. In this work, we investigate for the first time the use of Curriculum Learning (CL) [BLCW09] in FL to boost the classification performance while improving domain alignment. Our CL approach is implemented via a data scheduler, which establishes a prioritization of the training samples. We assign higher importance to samples that are forgotten after the deployment of the global model. We show that presenting the training samples in this order is beneficial for FL, and also boosts the domain alignment between domain pairs. Similar to [LGD+20] we employ federated adversarial learning [PHZS19, PKM19] to deal with the alignment between the different domains. However, unlike Li *et al.* [LGD+20] that analyze 1-D signals extracted from f-MRI, we study the screening of high-resolution mammograms and use CL to boost the classification performance. We validate our strategy on a setup composed of one public and two private clinical datasets with non-IID intensity distributions. Different from [RCS+20], who proposes a FL framework for breast density classification and do not correct the misalignment between the domains, we target the more complex task of breast cancer classification. Furthermore, we propose a novel curriculum for the FL setting, and explicitly handle domain shift with federated adversarial domain adaptation.

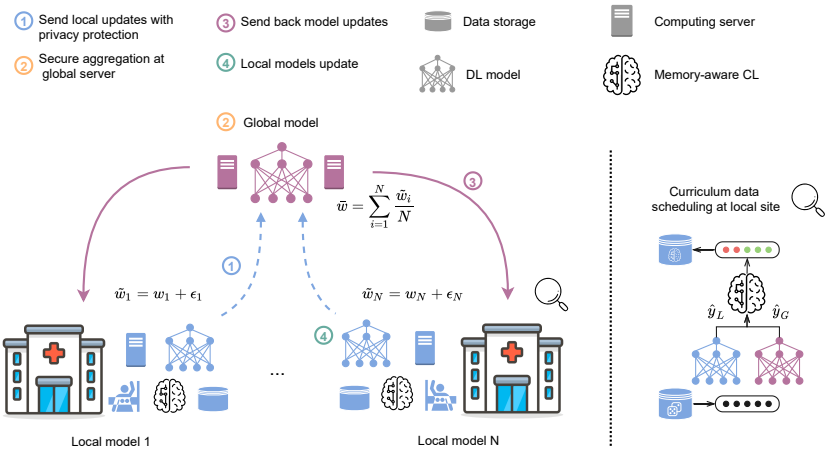

Figure 6.2. Memory-aware curriculum federated learning framework with data privacy protection. (1) Local models share their weights after the addition of Gaussian noise (dotted blue arrows). (2) The global server performs the aggregation of the local models' weights. (3) The resulting averaged model is deployed to each site (purple arrows). (4) Local models are updated. The curriculum data scheduler rearranges the training samples to prioritize samples that were forgotten after the deployment of the global model.

6.2 Related work

6.2.1 Federated Learning

FL arises from the need of sharing sensitive medical data between different healthcare providers. FL has been mainly formulated in two ways: (i) differential privacy [DKM+06, DR+14], *i.e.* each site trains a local model with private data and only shares model parameters [ZNH+18], and (ii) protecting the details of the data using cryptographic techniques [BIK+17], such as secure multi-party computation [MR18] and homomorphic encryption [HHIL+17]. We focus on the differential privacy approach.

Only few FL works have been shown effective on medical im-

ages. For instance, for brain tumor segmentation [SRE+18, LMX+19, BWRA21]; for prediction of disease incidence, patient response to treatment, and other healthcare events [HSQ+19]; and lately for classification [GHJK20, AdTBT20, LGD+20, WLD20, YFNA20]. Regarding breast imaging, only Roth *et al.* [RCS+20] have investigated breast density classification. As in [RCS+20], we employ a client-server-based FL method with `Federated Averaging (FedAvg)` [MMR+17b], which combines local Stochastic Gradient Descent (SGD) on each site with a server that performs model averaging. However, [RCS+20] significantly down-sampled the input mammograms. Although low resolutions are acceptable for density classification, the detail loss penalizes the malignancy classification task. Moreover, [RCS+20] did not apply any domain adaptation technique to compensate the domain shift of the different pixel-intensity distributions. Here, we opt for a different approach by working on high-resolution mammograms with federated domain adversarial learning [PHZS19].

6.2.2 Domain Adaptation

Deep learning methods assume that samples from the training (source) and testing (target) set are IID data. However, this statement does not always hold. When the data distribution from the source and target domains is related but different, there is a domain shift. Domain Adaptation (DA) aims to remove such shifts by transferring the learned representation from a source to a target domain. When target labels are unavailable during the training phase, UDA techniques are employed. One of the UDA strategies is to learn a domain-invariant feature extractor, which aligns the feature distribution of the target domain to that of the source by: (i) minimizing a distance of domain discrepancy [LWD+13], (ii) revisiting batch normalization layers [CPC+17], or (iii) through adversarial learning [GUA+16].

Despite less annotation requirements, the above UDA approaches need access to both source and target data [LCWJ15b, GL15]. However, in the federated setting, data is stored locally and cannot be

shared. Recently, federated batch normalization [LJZ+21] and federated adversarial domain adaptation [PHZS19, PKM19] have been proposed to deal with DA under the privacy-preserving requirement. The work by Li *et al.* [LJZ+21] focus on mitigating *feature shift*, *i.e.* the deviation in feature space, using batch normalization before averaging the local models. Whereas Peng *et al.* [PHZS19] train in an adversarial manner a feature extractor and a domain discriminator to learn a domain-invariant representation and alleviate domain shift. The latter has been applied to f-MRI on 1-D signal data using a multi-layer perceptron [LGD+20]. Different from the work by Li *et al.* [LGD+20], we study federated adversarial alignment using a deep convolutional neural network (ResNet-22) on medical images, in particular, high-resolution mammograms for breast cancer classification.

6.2.3 Curriculum Learning

CL [BLCW09] is inspired in the *starting small* concept from cognitive science. CL methods follow a systematic and gradual way of learning. A scoring function is defined to determine the priority of the training samples. Based on this scoring function, which can measure, for example, difficulty or uncertainty, the training samples are weighted or presented in a certain order to the optimizer. This new order has an impact on the local minimum achieved by the optimizer, leading to an improvement in the classification accuracy.

CL has already demonstrated an improved performance in medical image classification tasks, such as thoracic disease [TWH+18], skin disease [YWL+19], proximal femur fractures [JSMK+19**?**] and breast screening classification [MBN+18]. These techniques exploit either attention mechanisms [TWH+18], meta-learning [MBN+18], prior knowledge [JSMK+19, YWL+19] or uncertainty in the model's predictions [JSMK+21].

There is little prior work in CL in combination with DA techniques for general classification. Mancini *et al.* [MARC20] investigated a combination of CL and Mixup [ZCDLP17] for recognizing

unseen visual concepts in unseen domains. Shu *et al.* [SCLW19] addressed two entangled challenges of weakly-supervised DA: sample noise of the source domain, and distribution shift across domains. An extreme case of DA is that of zero-shot learning, in which at test time, a learner observes samples from classes that were not observed during training. Tang *et al.* [YBLS20] proposed an adversarial agent, referred to as curriculum manager, which learns a dynamic curriculum for source samples.

Different from [MARC20, ZCDLP17, SCLW19, YBLS20] that aim at improving transferability between domains, we choose to schedule the data within each domain. We design local data schedulers aiming to improve the consistency between global and local models and prevent forgetting samples that were previously correctly classified by the local model. To this end, we monitor the training samples before and after the deployment of the global model. We define a scoring function that assigns high values to samples that have been forgotten by the local model. Thus, our CL method builds locally memory-aware data schedulers to avoid forgetting.

6.3 Methods

In this section, we formulate the details of our proposed curriculum approach to locally schedule training samples in the FL setting. The overall FL framework is depicted in Fig. 6.2. In this setting, we assume that each local site has data storage, a computing server and a memory-aware CL module. Nevertheless, at the global level, no imaging data are stored and only computing is possible. In this type of FL setting, it is common to share the model weights and aggregate them at the central service. Moreover, local healthcare providers may have diverse imaging systems resulting in datasets with different intensity profiles. To ease the existing domain shift between the sites, we deploy an UDA strategy that shares the latent representations (and not the image data) between domain pairs. Both the model weights and

the embeddings are blurred with Gaussian noise [LGD+20] to protect the private data using differential privacy [DKM+06, DR+14]. The memory-aware CL module compares the local and global model predictions and assigns scores to each training sample. The data scheduler leverages the curriculum probabilities to locally arrange the samples.

In Subsection 6.3.1, the overall FL framework is presented. In Subsection 6.3.2, we present the details of the FL setup with data privacy-preserving scheme. Then, in Subsection 6.3.3, we introduce DA into the framework. And finally, in Subsection 6.3.4 we present the details of our proposed method leveraging CL to avoid forgetting locally learned samples in the FL setting.

6.3.1 Multi-site learning

Next, we develop our method to learn a collaborative CAD system in a decentralized multi-site scenario with a privacy-preserving strategy. Let us denote each site's dataset as $\mathcal{D}_n$ where $n = 1, ..., N$ and N is the total number of sites. Each dataset is composed of mammography images X_n and their corresponding diagnosis Y_n, *i.e.* $\mathcal{D}_n = \{X_n, Y_n\}$. We aim to detect malignant cases by training a deep-learning model. We formulate the learning objective as a binary classification task, where malignant samples correspond to the positive class. Each local model aims to minimize the cross-entropy loss over the training data from a particular site n:

$$\mathcal{L}_{Cls,n} = -\sum_{n_k} y_{n_k} \log(p_{n_k}) + (1 - y_{n_k}) \log(1 - p_{n_k}), \tag{6.1}$$

where y_{n_k} is the label of the k-th subject in the training label set $Y_n = \{y_{n_1}, ...y_{n_{|Y_n|}}\}$ and p_{n_k} is the corresponding output probability of the model for an input $x_{n_k} \in X_n$. As depicted in Fig. 6.3-*left*, we split the deep learning model into a feature extractor F and a classifier Cls. We refer to the output of the feature extractor as the latent representation or embedding. In this work, we assume the

most challenging scenario, in which we consider that each site has mammography systems of different vendors (see Fig. 6.1).

6.3.2 Federated learning

We assume that data owners collaboratively train a global model without sharing their image data. The term *federated* was coined because the learning task is solved by a federation of participating models (frequently referred to as *clients*), which are coordinated by a central *server*.

The FL scenario is depicted in Fig. 6.2. We assume that each local site has data storage and a computing node. Nevertheless, at the global level, only computing is possible. Once that individual models have been trained on private data, there are four key steps in the FL training process: (1) local updates are sent to the global server with privacy protection or encryption, (2) the central server aggregates the local updates, (3) the aggregated model parameters are deployed to the local sites, and (4) local models are updated. After that, a new round of local training starts.

To apply SGD in the federated setting, each client n computes gradients on the full local data for the current model, and the central server performs the aggregation of these weights to build a global update. Let us assume a fixed learning rate η and denote the gradients at each client as g_n. The central server computes the update as $w_{t+1} \leftarrow w_t - \eta \sum_{n=1}^{N} \frac{m_n}{M} g_n$, where m_n is the number of images at site n, and M the total number of images. We refer to this algorithm as `FedSGD`. We can decompose the global update into local client ones: first, one takes a gradient descent step from the current model using each local dataset, $\forall n \, w_{t+1}^n \leftarrow w_t - \eta \, g_n$. Then, we let the server make a weighted average of the resulting local updates as $w_{t+1} \leftarrow \sum_{n=1}^{N} \frac{m_n}{M} w_{t+1}^n$. Instead of performing one global update after each local computation, we can add multiple iterations of the local update to each client before the averaging step. Model updates are performed at every *communication round*. Let us denote: Q,

the total number of optimization iterations; τ, the communication pace; and B, the local mini-batch size used for the client updates. In each epoch, the communication between the models happens Q/τ times. Federated averaging `FedAvg` [MMR+17b] is a generalization of `FedSGD`, which allows local nodes to perform more than one batch update on local data and exchanges the updated weights rather than the gradients. We build on top of `FedAvg` to further consider domain alignment.

6.3.3 Federated adversarial learning

Medical images collected from different healthcare providers may originate from diverse devices or imaging protocols, leading to non-IID pixel-intensity distributions. In this scenario, we try to compensate the domain shift between every pair of domains. There is extensive literature on UDA methods [GL15, HMZ18, ZPIE17]. However, these works do not generally satisfy the conditions of a FL setting: namely that data should be stored locally and not shared. To satisfy the requirements of the FL framework and to address the domain shift problem, we rely on federated adversarial alignment [PHZS19]. This method aligns the feature space by progressively reducing the domain shift between every pair of sites. To preserve privacy, only the noisy latent representations (Gaussian noise is added to each local latent representation) are shared between the sites every communication round. This method leverages a domain-specific local feature extractor F, and a global discriminator D. For source $\mathcal{D}^S$ and target $\mathcal{D}^T$ sites, we train individual local feature extractors F^S and F^T, respectively. For each $(\mathcal{D}^S, \mathcal{D}^T)$ source-target domain pair, we train a domain discriminator D to align the distributions.

Optimization takes place in two iterative steps. In the first, the objective for discriminating the source domain from the others is

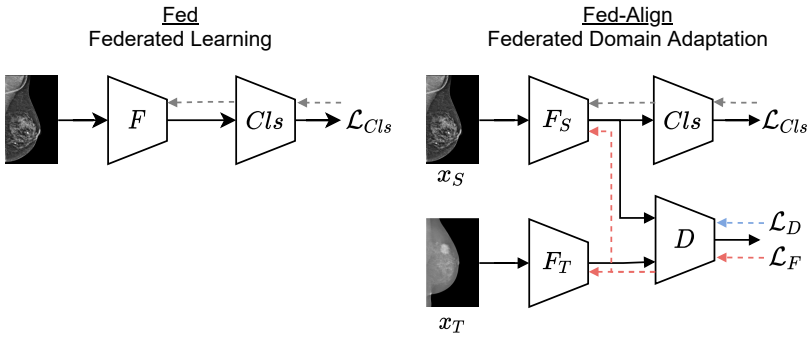

Figure 6.3. Architecture comparison of *left*: Fed and *right*: Fed-Align. Colour dotted lines indicate backward passes with respect to each loss function. ($\mathcal{L}_{Cls}$: Eq. (6.1), $\mathcal{L}_D$: Eq. (6.2), $\mathcal{L}_F$: Eq. (6.3), F: feature extractor, Cls: classifier, D: domain discriminator).

defined as:

$$\begin{aligned}\mathcal{L}_D(X^S, X^T, F^S, F^T) = &- \mathop{\mathbb{E}}_{x^S \sim X^S}[\log D\left(F^S(x^S)\right)] \\ &- \mathop{\mathbb{E}}_{x^T \sim X^T}[\log\left(1 - D\left(Z \circ F^T(x^T)\right)\right)],\end{aligned} \tag{6.2}$$

where $Z(\cdot)$ is the Gaussian noise generator for privacy preservation. In the second step, we consider the adversarial feature extractor loss:

$$\begin{aligned}\mathcal{L}_F(X^S, X^T, D) = &- \mathop{\mathbb{E}}_{x^S \sim X^S}[\log D\left(F^S(x^S)\right)] \\ &- \mathop{\mathbb{E}}_{x^T \sim X^T}[\log(D\left(Z \circ F^T(x^T)\right)].\end{aligned} \tag{6.3}$$

The weights of the feature extractor F and the domain discriminator D remain unchanged for the first and second step, respectively.

6.3.4 Memory-aware curriculum federated learning

We propose to incorporate CL to improve the classification performance of the federated adversarial learning approach. In particular,

the implementation of the curriculum is in the form of a data scheduler. A data scheduler is a mechanism that controls the order and pace of the training samples presented to the optimizer. We follow our previous work [**?**] and tailor it for the federated setting. In the following, we introduce the required components to define our CL method. We formalize the definition of the data scheduler through three components: a scoring function ρ, curriculum probabilities γ, and a permutation function π, and provide further details in the next paragraph.

The key element of our approach is the scoring function ρ, which is specific for FL. The scoring function ρ assigns a score to every sample, which normalized becomes a curriculum probability γ. These probabilities are then used to sample the training set $\{X, Y\}$. The sampling operation establishes a permutation π determining the reordered dataset $\{X_\pi, Y_\pi\}$, finally fed in mini-batches to the optimizer.

We consider a dynamic approach in which the scoring values are computed at every epoch e for every training sample k. We get the predictions at every site n, before and after the communication between the models, obtaining local and global predictions $\hat{y}_L$, $\hat{y}_G$, respectively. To avoid forgetting in the FL setting, our scoring function ρ assigns higher values (thus higher curriculum probabilities γ) to samples that were forgotten. The order in which samples are presented to the optimizer is determined by the curriculum probabilities γ. Our function is defined as:

$$\rho_k^{(e)} = \begin{cases} 2.0 & \text{if } \hat{y}_{L,k}^{(e)} = y_k \text{ and } \hat{y}_{G,k}^{(e)} \neq y_k \\ 1.0 & \text{otherwise.} \end{cases} \tag{6.4}$$

We emphasize learning of samples for which the prediction changed from correct to wrong after the model aggregation. Our memory-aware curriculum federated learning method is summarized in Algorithm 1 (Suppl. Material).

6.4 Experimental validation

In order to validate the effect of data scheduling on the breast cancer classification, we perform experiments with two private and one public dataset. We compare our proposed approach combining FL, DA and CL, against FL alone and FL with DA.

6.4.1 Datasets

For our study, we employ 3 datasets of Full Field Digital Mammography (FFDM), coming from three different vendors: Hologic, GE and Siemens (INBreast [MAD+12]). The first two are private clinical datasets, and the last one is publicly available. Institutional board approvals were obtained for each of the datasets. Intensity profiles [SMR+19] among the datasets varied significantly, and can also be observed in Fig. 6.1. This variability is mainly due to the different mammography systems and acquisition protocols used to generate digital mammograms. We do not use any site-specific image filtering to compensate the domain shift and we apply the same preprocessing to the images from the three sites. The preprocessing consists of standard normalization with mean subtraction and division by the standard deviation. Each dataset was split into three parts with the ratio approximately of 70%:10%:20% to build respectively the training, validation and test sets. Our problem is formulated as a binary classification task. The number of samples per class and database can be found in Table 6.1. The first class reunites benign findings and normal cases, the second class contains only malignant cases confirmed with a biopsy. Mammography images are of different sizes, we cropped the empty rows and columns, and resized to 2048 pixels in height, and then padded to 2048 pixels in width. It is often the case that important cues for diagnosis are subtle findings in the image, which could be as small as 10 pixels in length [Mer14]. Therefore, we do not apply any further downsampling and use a resolution of 2048 pixels, close to the original resolution.

	Hologic	Siemens	GE
Total Subjects	1460	410	852
Benign/Normal	730	287	421
Malignant	730	123	431

Table 6.1. Summary of the datasets used in this study.

6.4.2 Experimental Setting

We perform an in-depth evaluation of our proposed method *Fed-Align-CL* with a series of experiments. First, we investigate the effect of different pretraining strategies in the FL framework. Second, we compare the classification performance of our approach against other non-federated and federated approaches. Third, we investigate the influence of DA and CL in the resulting feature embeddings of the different methods.

6.4.3 Implementation details

Architectures. We employ as feature extractor F the architecture proposed by Wu *et al.* [WPP+19], a ResNet-22 [HZRS16a] that is adapted to take high-resolution images ($\sim$ 4 megapixels) as input. We initialize the feature extractor with the pretrained weights provided by Wu *et al.* [WPP+19][1]. The weights of the classifier Cls and domain discriminator D are randomly initialized. The classifier Cls is formed by 3 fully-connected layers. The first two are followed by batch normalization, ReLu activation, and dropout. The architecture for the domain discriminator D is formed by two fully-connected layers with a ReLu activation in between and a sigmoid layer for the final output. Details of the architecture of the models can be found in Table 6.4 of the Supplementary Material. Our memory-aware CL approach builds on top of the federated adversarial learning code pro-

[1] https://github.com/nyukat/breast_cancer_classifier

vided by Li *et al.* [LGD+20][2]. Different from [LGD+20] that employs a multi-layer perceptron for 1-D f-MRI signals, we deploy a specific CNN for high-resolution mammography images.

Hyperparameters. We train our models 5 times with different seed initialization for the classifier Cls and domain discriminator D. Adam optimization is used for 50 epochs with an initial learning rate of 1e−5. We compute the adversarial domain loss L_D, and also introduce the curriculum data scheduling, after training the feature extractor F and classifier Cls for 5 epochs. The dropout rate for the classifier Cls is set to 0.5. The number of optimization iterations $Q = 120$, and the local batch size $B_n = \lfloor m_n/Q \rfloor$. In each epoch, local models are updated according to the communication pace τ. The shared weights are modified by the addition of random noise ϵ to protect data from inverse interpretation leakage. We generated Gaussian noise $\epsilon \sim N(0, s_h^2\sigma^2)$, assuming a sensitivity $s_h = 1$ and a variance $\sigma^2 = 0.001$. We investigated different communication paces $\tau = \{10, 20, 40, 60\}$, and noise values $\sigma^2 = \{0, 0.001, 0.01, 0.1\}$. We did not find significant differences in classification accuracy for the different communication paces τ. There is a direct correlation between the amount of noise introduced in the system and the model performance, we consider that adding a noise $\sigma^2 = 0.001$ is a good trade-off.

Evaluation metrics. For the classification task, we report the area under the receiver operating characteristic curve (ROC-AUC) and AUC for the precision-recall curve (PR-AUC).

6.5 Results

Initialization of local models. First of all, we investigate the classification performance of the FL method with different pretrain-

[2]https://github.com/xxlya/Fed_ABIDE

Initialization	**Hologic**	**Siemens**	**GE**	**AVG**
Local model	0.57	0.38	0.66	0.53
Scratch	0.73	0.52	0.65	0.63
DDSM	0.69	0.62	0.65	0.65
Wu *et al.* [WPP+19]	**0.78**	**0.65**	**0.83**	**0.75**

Table 6.2. AUC of the federated learning method using different initialization approaches.

	Hologic		**Siemens**		**GE**		**AVG**	
	AUC	PR-AUC	AUC	PR-AUC	AUC	PR-AUC	AUC	PR-AUC
Wu *et al.* [WPP+19]	0.65	0.69	0.67	0.75	0.79	0.78	0.70	0.73
trHologic	-	-	0.67	0.74	0.73	0.74	-	-
trSiemens	0.59	0.63	-	-	0.65	0.67	-	-
trGE	0.64	0.66	0.72	0.79	-	-	-	-
Single	0.83	0.84	0.83	0.84	0.85	0.83	-	-
Fed	0.78	0.78	0.65	0.74	0.83	**0.83**	0.75	0.77
Fed-CL	0.80	0.80	0.63	0.72	0.81	0.81	0.75	0.78
Fed-Align	0.79	0.78	0.69	**0.79**	**0.85**	**0.83**	0.78	0.80
Fed-Align-CL	**0.84**	**0.84**	**0.70**	**0.79**	0.83	0.82	**0.79**	**0.82**
Mix	0.83	0.84	0.86	0.83	0.82	0.88	0.84	0.85

Table 6.3. Comparison of strategies. Median AUC and PR-AUC of the 5 runs, except for Wu *et al.* [WPP+19]. The highlighted values in bold correspond to the best federated method.

ing strategies. The first case *Local model* corresponds to pretraining each model with their own private data. The second case *Scratch* to a random initialization of the local model weights. The third case *DDSM* corresponds to pretraining the models on the CBIS-DDSM dataset [LGH+17]. The last case corresponds to initializing the model with the publicly shared weights from Wu *et al.* [WPP+19].

In Table 6.2, the AUC for the different initialization strategies is reported. We found that the best approach was using the pretrained weights from Wu *et al.* [WPP+19]. This behaviour is expected because their model was trained with a very-large private

dataset. Moreover, the model in [WPP+19] was already pretrained with the ImageNet [KSH12a] dataset. Interestingly as well, classification results were better when all local models were initialized either randomly or pretrained on a single dataset (DDSM) than when each of them was pretrained on a small private dataset. Although DDSM dataset is large, it is formed by screen film mammography instead of FFDM, which explains the difference to Wu *et al.*'s weights.

Comparison with different strategies. To demonstrate the performance of our proposed method (*Fed-Align-CL*), three non-federated strategies and with two federated strategies. The non-federated strategies consists of: (i) training and testing within a single site (*Single*); (ii) training using one site and testing on another site (*Cross*); and (iii) collecting multi-site data together for training (*Mix*). The later does not preserve data privacy since this model requires access to all training images and their respective classification labels. In *Cross*, we denote the site used for training as 'tr<site>'. Also, we ignore the performance of the site used for training in this row, and report it in the row 'Single'. The federated strategies consist of training a client-server-based FL method with: (iv) `FedAvg` [MMR+17b], and (v) federated adversarial learning [PHZS19, LGD+20]. We also performed an ablation study to verify the individual contributions of the domain alignment and the curriculum scheduling. Therefore, we included in our comparison *Fed-CL*.

Classification metrics for breast malignancy classification are reported in Table 6.3. In the first row, we include the performance of Wu's model [WPP+19] without further training. In the first place, we present the results of the non-federated methods. First, as expected, we find that the *Cross* models do not generalize well across manufacturers. Second, the individual models (*Single*) achieve an average AUC of 0.83 for the three sites. When comparing our performance to other works on the publicly available INBreast dataset (Siemens), we achieve an AUC comparable to [WPP+19], but lower than [RHU+18, SMR+19] with an AUC of 0.95. However, the later

two works rely on region-wise ground truth: the first leveraging ROI localization and the second one using patch pretraining. In contrast, our models only rely on the full mammograms and their corresponding classification label. As expected, the best performing model is *Mix*, which is trained with mammography images and their corresponding annotations from all sites, thus, not preserving privacy.

In the second place, we compare the federated approaches. First, we find that the *Fed-CL* method improves on average the PR-AUC with respect to *Fed*. However, the performance for the different domains of these two methods can be uneven. The *Fed-Align* approach helps to learn domain-invariant features that are beneficial for the classification task. Finally, we can see that our proposed *Fed-Align-CL* achieves on average the highest AUC and PR-AUC. We also find a consistent improvement with our proposed method when the models are trained from scratch. However, the classification metrics were better with the pretrained weights [WPP+19], as discussed in the previous experiment.

Alignment of features in latent space. In order to visualize the effect of the domain adaptation and curriculum scheduling techniques, we show in Fig. 6.4 the two-dimensional t-SNE [MH08] projection of the embedded latent space. First, we can see that the features learned by *Fed* are better clusterized according to the input domain, *i.e. Fed* learns domain-variant features. Second, the combination of FL with CL results in samples more spread along the manifold, although still dependent on the domain. In the plots that correspond to the models that perform DA, we can see that the input images are better clusterized according to the label instead of the domain. We find that the domain alignment is particularly helpful for Siemens. Fig. 6.5 in Supplementary Material shows a similar behaviour for the penultimate classification layer.

6.6 Discussion

FL is a potential solution for the future of digital health [RHL+20a], especially for classification tasks without access to sufficient data. FL allows for collaboratively training a model without sharing private data from different sites. A challenging aspect of sharing data within FL is that related to legal regulations and ethics. Privacy and data-protection need to be taken carefully into account. High-quality anonymization from a mammography or electronic record has to be guaranteed and in certain regions GDPR[1] [HM15] or HIPAA[2] [Ann03] compliant. Privacy-preserving techniques for FL provide a trade-off between model performance and reidentification. However, remaining data elements may allow for patient reidentification [RHdM19]. Unless the anonymization process destroys the data fidelity, patient reidentification or information leakage cannot be discarded.

Another challenging aspect of FL is that of training a model mixing heterogeneous data, *i.e.* images obtained from different system vendors or acquisition protocols. In this work, we have investigated and confirmed the negative effect of domain shift for the malignancy classification with multi-site mammograms. Models trained on single-vendor images did not generalize adequately to others. The best performing single-vendor model was the one for GE. Interestingly, we found that our curriculum federated learning approach did not improve in this case. This behaviour could be related to GE dataset being more similar to the pretrained weights of [WPP+19]. Moreover, due to the presence of domain shift between the datasets, the federated models that did not consider any domain adaptation performed worse than those that included domain alignment. We attribute this underperformance partially to the difference in the intensity profiles, and partially to the sizes of the datasets being insufficient for good generalization.

In this work, we have investigated the use of CL to boost the

[1]GDPR: EU/UK General Data Protection Regulation

[2]HIPAA: Health Insurance Portability and Accountability Act

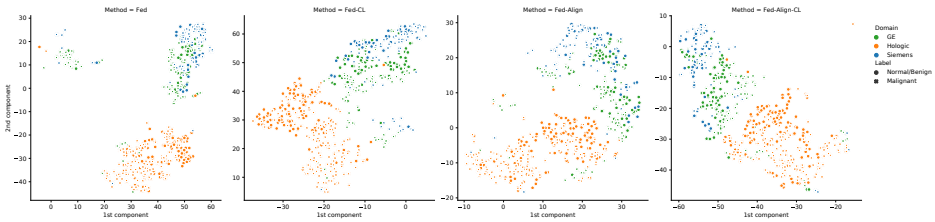

Figure 6.4. t-SNE visualization of the latent space obtained by Fed, Fed-CL, Fed-Align and Fed-Align-CL in that order. The circles represent normal and benign samples, and the crosses malignant cases. Each color represents a domain.

alignment between domain pairs and improve the overall classification of breast cancer. In particular, our memory-aware curriculum is implemented with a data scheduler that arranges the order of the training samples. This order is defined with a scoring function that prioritizes training samples that have been forgotten after the deployment of the global model. We believe further research will follow on the use of CL in combination with FL and DA. We envision three approaches: those focused on prioritizing training (source) samples for better classification; those focused on smartly weighting the aggregation of the local models; and those focused on improving alignment between domains pairs. Similar to the work presented in this study, other schemes can be designed to prioritize the (source) samples via a data scheduler, for instance, motivated by boosting [FSA99]. Regarding the local model aggregation, one could deploy a CL-based adaptive weighting for clients based on a dynamic scoring function taking into account meta-information [YFNA20], and in this way, help to cope with unbalanced and non-IID data. Finally, to improve alignment, scoring functions could rely on computing the distance between (noisy) latent representations of the source and the remaining domains to weigh each local model contribution.

6.7 Conclusions

In this work, we have designed and integrated a CL strategy in a federated adversarial learning setting for the classification of breast cancer. We have learned a collaborative decentralized model with three clinical datasets from different vendors. We have shown that, by monitoring the local and global classification predictions, we can schedule the training samples to boost the alignment between domain pairs and improve the classification performance.

Supplementary Material

Algorithm 1 presents the pseudo-code for our novel memory-aware curriculum federated learning.

Architecture of the models. We provide the detailed model architecture in Table 6.4. We denote convolutional layers as *Conv*, max pooling layers as *MaxPool*, fully connected layers as *FC*, batch normalization layers as *BN*, ReLu layers as *ReLu*, dropout layers as *Dropout* and sigmoid layers as *Sigmoid*. For FC layers, the values in brackets represent the input and output dimensions. For Conv layers, we provide in this order: the input and output feature maps, the kernel size, the stride and the padding. For MaxPool layers, we provide in this order: kernel size, stride, padding and dilation. For dropout layers (Dropout), we provide the probability of an element to be zeroed. To define the feature extractor, we define a *Block* which consists of a series of layers and specify for the convolutional layers: the input and output feature maps, the kernel size, the stride for the first convolution s_1, the stride for the second convolution s_2, and the padding. In particular Block consists of: ReLu, BN, Conv(), BN and Conv().

Statistical significance. In Table 6.5, we run a t-test between every pair of strategies to verify the significance of our results. We report the p-values between every federated strategy pairs.

t-SNE feature visualization. Figure 6.5 depicts the first two components after applying t-SNE to the penultimate classification layer of every federated method.

input : N number of sites, $X = \{X_1, ..., X_N\}$ mammograms, $Y = \{Y_1, ..., Y_N\}$ classification labels, $\{X_i^S, Y_i^S\}_{i=1}^N$ source dataset, $\{X_j^T\}_{j=1}^N$ target dataset, m_n training size at site n, $f_w = \{f_{w_1}, ...f_{w_N}\}$ local models, $Z(\cdot)$ noise generator for privacy-preserving, Q number of optimization iterations, τ communication pace, E optimization epochs, E_w warm-up epochs, $\{\text{opt}_1(\cdot), ..., \text{opt}_N(\cdot)\}$ optimizers

output: global model: g_w

Initialize local models: $\{f_{w_0}, ..., f_{w_N}\} \leftarrow$ Wu *et al.* [WPP+19]

for $e = 1$ **to** E **do**
  if $e > E_w$ **then**
    for $i = 1$ **to** N **do**
      Memory-aware curriculum
      for $k = 1$ **to** m_i **do**
        Get the local $\hat{y}_{L,k}^{(e)}$ and global $\hat{y}_{G,k}^{(e)}$ classification predictions ;
        Compute the curriculum weights $\rho_{i,k}^{(e)}$ with Eq. (6.4) ;
        Obtain reordering function $\pi_i^{(e)}$ by sampling with $\rho_i^{(e)}$;
        Reorder training data: $\{X, Y\} \xrightarrow[\pi_i^{(e)}]{} \{X_{\pi_i^{(e)}}, Y_{\pi_i^{(e)}}\}$;
      end
    end
  else
    Random permutation $\pi_i^{(e)}$;
  end

  for $q = 1$ **to** Q **do**
    for $i = 1$ **to** N **do**
      Local classification
      Get the **next** mini-batch from source site i $\{X_{\pi_i^{(e)},b}^S, Y_{\pi_i^{(e)},b}^S\}_{b=1}^{B_i * Q}$;
      Compute classification loss $\mathcal{L}_{Cls}(f_{w_i^{(q-1)}}(X_{\pi_i^{(e)},b}^S, Y_{\pi_i^{(e)},b}^S)))$ with Eq. (6.1) ;
      Update $w_{F_i}^{(q)}$, $w_{Cls_i}^{(q)} \leftarrow opt_i(\mathcal{L}_{Cls})$;

      if $e > E_w$ **then**
        for $j = 1$ **to** N ***and*** $j \neq i$ **do**
          Domain alignment
          Get the **next** mini-batch from target site j $\{X_{\pi_j^{(e)},b}^T\}_{b=1}^{B_j * Q}$;
          Compute adversarial loss $\mathcal{L}_D$ with Eq. (6.2) ;
          Update $w_{D_i}^{(q)} \leftarrow opt_i(\mathcal{L}_D)$;
          Compute feature extractor loss $\mathcal{L}_F$ with Eq. (6.3) ;
          Update $\{w_{F_i}^{(q)}, w_{F_j}^{(q)}\} \leftarrow opt_i(\mathcal{L}_F)$;
        end
      end

      if $q \,\%\, \tau = 0$ **then**
        Update global model
        $\bar{w}_F^{(q)} \leftarrow \frac{1}{N} \sum_{i=1}^N (w_{F_i}^{(q)} + Z(w_{F_i}^{(q)}))$;
        $\bar{w}_{Cls}^{(q)} \leftarrow \frac{1}{N} \sum_{i=1}^N (w_{Cls_i}^{(q)} + Z(w_{Cls_i}^{(q)}))$;
        Deploy weights to local model
        for $i = 1$ **to** N **do**
          $w_{F_i}^{(q)} \leftarrow \bar{w}_F^{(q)}$;
          $w_{Cls_i}^{(q)} \leftarrow \bar{w}_{Cls}^{(q)}$;
        end
      end
    end
  end
end

Return global model: $g_w \leftarrow (w_F^{(E)}, w_{Cls}^{(E)})$

Algorithm 1: Memory-aware Curriculum Federated Learning

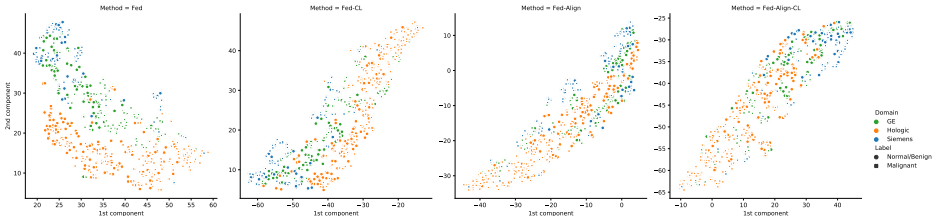

Figure 6.5. t-SNE visualization of the penultimate classification layer obtained by Fed, Fed-CL, Fed-Align and Fed-Align-CL in that order. The circles represent normal and benign samples, and the crosses malignant cases. Each color represents a domain.

Layer	Configuration
	F: Feature Extractor
1	Conv(1, 16, 3, 1, 1), MaxPool(3, 2, 0, 1)
2.1	Block(16, 16, 3, s_1=1, s_2=1, 1), Conv(16, 16, 1, 1)
2.2	Block(16, 32, 3, s_1=1, s_2=1, 1)
3.1	Block(16, 32, 3, s_1=2, s_2=1, 1), Conv(16, 32, 1, 2)
3.2	Block(16, 32, 3, s_1=1, s_2=1, 1)
4.1	Block(32, 64, 3, s_1=2, s_2=1, 1), Conv(32, 64, 1, 2)
4.2	Block(32, 64, 3, s_1=1, s_2=1, 1)
5.1	Block(64, 128, 3, s_1=2, s_2=1, 1), Conv(64, 128, 1, 2)
5.2	Block(64, 128, 3, s_1=1, s_2=1, 1)
6.1	Block(128, 256, 3, s_1=2, s_2=1, 1), Conv(128, 256, 1, 2)
6.2	Block(128, 256, 3, s_1=1, s_2=1, 1)
	Cls: Classifier
1	FC(256, 128), BN, ReLu, Dropout(0.5)
2	FC(128, 64), BN, ReLu, Dropout(0.5)
3	FC(64, 2), Sigmoid
	D: Domain Discriminator
1	FC(256, 4), ReLu
2	FC(4, 2), Sigmoid

Table 6.4. ResNet-22 architecture for breast cancer classification.

	Fed		Fed-CL		Fed-Align	
	AUC	PR-AUC	AUC	PR-AUC	AUC	PR-AUC
Fed-CL	$7.60E-01$	$7.13E-01$	-	-	-	-
Fed-Align	$2.34E-03$	$1.51E-03$	$1.55E-01$	$1.82E-01$	-	-
Fed-Align-CL	$1.46E-04$	$9.82E-06$	$5.75E-02$	$3.29E-02$	$5.65E-02$	$1.42E-02$

Table 6.5. P-values to test statistical significance of the 5 runs among the different methods.

7 Conclusions

7.1 Summary of findings

In this dissertation we investigated different architectures and training strategies to learn deep-based representations for medical image classification. We provided readers with a comparison of architectures (between capsule and convolutional neural networks), and a unified framework for scheduling data based on curriculum learning, also for the case of a federated learning scenario.

7.1.1 Architectural design

We investigated in Chapter 3 the role of a localization architecture prior to the classification with a CAD tool. In Chapter 4, we researched the use of an alternative representation (capsules) through a different optimization scheme (routing-by-agreement).

Importance of abnormality localization

In Chapter 3, we presented a supervised localization CNN for the detection of a ROI in X-ray images. The localization was formulated

as an auxiliary task prior to the classification of proximal femur fractures. The localization of the ROI was highly accurate, all the predicted centers of the ROI were contained in the manually provided bounding box. This work was done in the beginning of the thesis, one could explore more recent supervised approaches like transformers [CMS+20] or weakly supervised methods [RYY+20, HZKH20]. For our task, the localization of the femur, we found that formulating the localization as the regression of the bounding box' coordinates achieved satisfasctory results. We noticed that there was not a unique scale for the definition of the ROI, *i.e.* how much context around the fracture was required for a proper classification. Therefore, we analyzed in detail the sensitivity of the CAD system to the size of the ROI. We found that disagreement in classification at different ROI sizes could signal misclassified examples. This analysis was also carried out before the now standard test time augmentation.

Impact and adoption into clinical practice of a CAD system

Given the results of Chapter 3, we made a preliminary analysis on the clinical impact of a CAD tool for the classification of proximal femur fractures. In particular, we further discussed two possible ways to integrate a CAD system based on our method to assist trauma surgery residents (Section 3.6). First, we believe that such a fracture classification tool could be used as a second reading. The tool could help in reducing fatigue while improving accuracy, given that the machine predictions are not affected by experience or workload [Sum10]. Second, we showed that the CAD system can retrieve similar images to a query radiograph. This tool could provide support for ambiguos cases to reach a more adequate treatment decision. Both a second reading or the retrieval tool could be particularly useful to assist the training of trauma surgery residents, especially for those working in small peripheral hospitals [Doi07]. To further verify our hypothesis an usability study should be carried on.

Equivariance to deal with class-imbalance or limited annotations

In Chapter 4, we experimentally validated the effectiveness of using capsule networks to improve CAD classification performance under medical data challenges. In particular, we demonstrated the increased generalization ability of capsule networks *vs.* CNNs when dealing with limited amounts of data and class-imbalance. We considered in our study a total of four datasets: two common computer vision datasets: MNIST and Fashion-MNIST; and two medical datasets: one for mitosis detection, and a second one for the detection of diabetic retinopathy.

The performance improvement was a result of the ability of capsule networks to model equivariance, that is, its ability to learn pose parameters along with filter weights. Together with the *routing-by-agreement* algorithm, this paradigm change required to see fewer viewpoints of the object of interest, and therefore fewer images, in order to learn the discriminative features to classify them. We found that capsule networks without using data augmentation were able to achieve a similar or better classification performance than CNNs using data augmentation. These results confirmed the benefits of equivariance over invariance.

Limitations of capsule networks

We also reported limitations to this otherwise general improvement of capsule networks over CNNs. In particular, capsule's performance improvement was significant in many cases but had a limit for the more complex datasets. Also, classification tasks where the global spatial structure plays a role can better exploit the advantages of capsule networks. However, the routing-by-agreement algorithm processes images patch-wise, which can be suboptimal for some tasks. Routing-by-agreement is also slower than regular backpropagation. In addition, capsule networks lack purposed layers, e.g. batch normalization, that could help to ease the convergence. Finally, when

visualizing the images reconstructed through the encoder-decoder branch (Fig. 4.3), we observed that they were blurry, especially for medical datasets with complex backgrounds. The fully-connected layers of this branch seemed to be good enough to regularize the parameter optimization but lost a considerable amount of information.

7.1.2 Training design

We presented in Chapter 5 a CL-inspired framework to schedule the order and pace of training data. In Chapter 6, we dealt with a more challenging scenario: collaboratively training a model without sharing private non-IID data from different sites.

Leveraging medical knowledge to guide learning

In Chapter 5, we addressed the challenging task of fine-grained proximal femur classification according to the AO standard. This task is especially hard due to the high intra- and inter-expert disagreement [vERMR10]. The assessment of medical images and their annotation is often subjective, therefore, disagreement in the annotations is frequent on medical image datasets. Our results confirmed that the information derived from medical guidelines, decision trees or inconsistencies in the annotations are effective and compatible with modern DL approaches.

Scheduling data with prior knowledge and uncertainty

We identified in Chapter 5 common scheduling elements for multi-class classification, and presented them in a unified CL framework. The definition of the CL scheme was based on a scoring function, that assigns a rank to each training sample, and a pacing function, that schedules the proportion of the training data to be used. Both functions are updated at every epoch. In our framework, we proposed three variants based on the scoring and pacing functions. The first one consists of a reordering of the whole training set, the second one

proposes sampling a training subset, and the third one assigns weights per training sample, following the same principle that a boosting strategy [FSST97].

To prioritize training data, we proposed two types of ranking functions leveraging: prior knowledge and uncertainty. We validated the benefits of our approach for the classification of proximal femur fractures based on the AO standard, and reached a performance comparable to state-of-the-art and experienced trauma surgeons. Furthermore, in controlled experiments with the MNIST dataset, we showed that the proposed method is effective for datasets with class-imbalance, limited or noisy annotations. From our experiments, we concluded that for datasets of limited size or under the presence of class-imbalance, the use of the subsets variant leads to an improved classification performance. One can either exploit prior knowledge to achieve a better performance, or if the computational cost is not an issue, leverage uncertainty. In the case of unreliable labels, we found that the more advantageous approach was the combination of weights with uncertainty.

Memory-aware curriculum federated learning

Our findings in Chapter 6 reinforce the importance of collaborative learning, and highlight the relevance of DA in this setting. FL has been highlighted as a potential solution for the future of digital health [RHL+20a], especially for classification tasks without access to sufficient data. The reason behind is that FL allows for collaboratively training a model without sharing private data from different sites.

Training a model in a collaborative manner needs to address three main challenges: privacy preservation, datasets with different pixel-intensity distributions and distributed optimization. We dealt with these challenges in the context of FL for breast cancer classification. We investigated the use of CL to improve the overall classification of breast cancer classification. Our approach was combined with UDA to deal with domain shift while preserving data privacy. In particular,

our memory-aware curriculum was implemented with a data scheduler that arranges the order of the training samples. This order was defined with a scoring function that prioritizes training samples that were forgotten after the intermediate updates with the global model. We evaluated and verified the effectiveness of our curriculum federated learning approach for the classification of breast cancer from mammograms on a collaborative setup with three clinical datasets from different manufacturers (two private and an open database).

7.2 Future work

Based on the findings of this thesis, we discuss next open research challenges that we have identified and that would be interesting to address in the future.

7.2.1 Curriculum for dynamic routing

In Chapter 4, we presented an extensive evaluation of the use of capsule networks. In this study, we employed two computer vision datasets for multi-class classification, and two medical datasets for binary classification. There have been many follow-up works regarding capsule networks and dynamic routing [HSF18, LB18b, KSTH19, STY+20]. To overcome the limitations of the input size and local connectivity, LaLonde *et al.* [LB18b] proposed a *locally-constrained dynamic routing*, which operates on large images (512×512 pixels) and employs capsules in a U-Net fashion. In [HSF18], the encoding of the pose is upgraded. Each capsule layer corresponds to a Gaussian whose mean is the pose. A pose matrix is used instead of a vector, and dynamic routing is updated to Expectation-Maximization routing. In [KSTH19] an unsupervised version of capsule networks called stacked capsule autoencoders is proposed. Unlike the original capsule network [SFH17], stacked capsule autoencoder is a generative model with an affine-aware decoder. This forces the encoder to learn an image representation that is equivariant to viewpoint changes. Also

in an unsupervised manner, motion is exploited as a powerful perceptual cue for part definition in [STY+20]. Learning with FlowCapsules is accomplished using flow estimation from capsule shapes and poses as a proxy task. Given the nature of part-whole relationships within capsule networks, the integration of CL to gradually discover more complex relationships could be an interesting area of future research. For instance, a CL approach could be formulated with a data scheduler controlling the order and pace of training samples presented to the model.

7.2.2 Uncertainty and evidential theory

The sources of uncertainty in CNNs largely fall into two categories: epistemic or model uncertainty and aleatoric or data uncertainty. Epistemic uncertainty refers to uncertainty caused by a lack of knowledge (about the best model). Aleatoric uncertainty refers to the notion of randomness, *i.e.* the variability in the outcome of an experiment that is due to inherently random effects. Because of their nature, epistemic uncertainty refers to the reducible part of the (total) uncertainty, whereas aleatoric refers to the irreducible part. In our proposed CL framework (Chapter 5), we extracted uncertainty from model's predictions to define the scoring function of the CL method. We restricted our study to the predictive entropy of the model, which includes both aleatoric and epistemic uncertainty. We reckon that assessing separately each type of uncertainty could be advantageous for some applications. Samples' uncertainty was measured with MC dropout. Instead, one could investigate the use of a Dirichtlet distribution to parametrize the output of the network. Then, the behavior of such predictor could be interpreted from an evidential reasoning perspective, such as in subjective logic [SKK18, Jøs18].

7.2.3 How to define a curriculum

CL aims at making the cost function easier to minimize by increasing the influence of simpler examples. In the work presented in this thesis, we showed that this can be achieved by sampling easier samples more frequently or by assigning them larger weights in the new cost function. We showed that, it is possible and benefitial, to integrate knowledge extracted from medical guidelines or decision trees to improve the classification task. We proposed this integration through data schedulers. The data scheduler employs a scoring function that ranks the priority of each training sample. The scores of the training samples were either defined a priori or they were dynamically estimated based on the model's performance. Depending on the classification scenario, defining the scoring function at the sample or class level could be better. We discussed the advantages and disadvantages of some cases, for example, we found more effective defining the score function at the sample level when noisy annotations were present in the dataset.

Other possibilities to ease the objective function could involve the addition of extra regularization terms (constrained optimization) or auxiliary tasks. Regarding the first aspect, Kumar *et al.* [KPK10] proposed to jointly learn the CNN's weights and sampling weights by including a regularization term into the objective function. Building upon this idea, recent works focus on boosting the diversity of the training data [JMY+14, ZB18, STD19]. The later one is particularly interesting because also differentiates between class- and instance-level for the curriculum definition. In this method, the *data parameters* leverage Knowledge Distillation [HVD15] to weigh the training logits, *i.e.* these parameters act as temperature. In our study, we restricted our methods to CL-based data schedulers, exploring the definition of auxiliary tasks with a Teacher-Student setup or constrained optimization are interesting research directions.

7.2.4 Curriculum for model aggregation in federated learning

Most works related to collaborative learning employ a client-server-based FL method with `Federated Averaging (FedAvg)` [MMR+17b], which combines local Stochastic Gradient Descent (SGD) on each site with a server that performs model averaging. However, equally weigthing the local models may not be optimal when there exists statistical heterogeneity in data. Recently, alternative aggregation strategies have been proposed [YFNA20, GHJK20]. Yeganeh *et al.* [YFNA20] proposed inverse distance aggregation with the objective of handling unbalanced and non-IID data for the classification of dermatoscopic images. Inverse distance aggregation proposes to adaptively weight the contribution of the clients on the inverse distance of each client parameters to the average model of all clients. Grimberg *et al.* [GHJK20] presented weight erosion for the task of survival prediction. This method is conceptually related to local fine-tuning. The difference is that the weight erosion scheme is optimized to discard contributions from unhelpful clients as early as possible in the training process.

Similar to [YFNA20, GHJK20], we could leverage CL for the model aggregation in FL. Instead of defining the scoring and pacing functions for the individual training samples, we could define them for each site. Then, one would need to define a measurement to prioritize the order or weight given to the local sites. One could also choose to include all models in the aggregation or not. Furthermore, instead of establishing a priori the pace for the communication rounds (between local model and global server), the communication pace could be derived from local or global model's performance.

7.2.5 Knowledge distillation for federated learning

Knowledge distillation [HVD15] was proposed as a compression technique: to transfer the knowledge from a cumbersome model to a lighter model. The model providing knowledge is referred to as *Teacher* and the model learning the knowledge is referred to as *Student*. Knowledge distillation is implemented by introducing a *temperature* factor to obtain weighted logits, referred to as "soft targets". When the soft targets have high entropy, they provide more information per training case than hard targets and much less variance in the gradient between training cases. Therefore, knowledge distillation can help to train models on much less data [HVD15] and to overcome overfitting on corrupted labels [JZL+18]. The federated setting resembles that of *Student-Teacher* learning. KD could be used to compress the local models that are sent to the global server. CL could be integrated in this framework to control the pace of compression or the weights for aggregation of the model.

7.3 Final remark

We have enumerated in the previous section numerous future work directions related to this doctoral thesis in the lines of capsule networks, curriculum strategies and FL. Our work payed special attention to the challenging aspects of medical image datasets: noisy labels, domain shift, class-imbalance and limited data. The clinical adoption of DL will also depend on handling conditional challenges: bias, interpretability, transparency, *etc.*

A

Introduction to Convolutional Neural Networks

A.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) [ON15] are computational processing systems heavily inspired by how biological nervous systems operate. These networks are composed of a high number of interconnected computational nodes, frequently referred to as neurons, which work in an entwine fashion to collectively learn to optimize an expected output (from the input).

The basic structure of an ANN is composed of an input layer, some hidden layers and an output layer, as depicted in Fig. A.1. The input is usually a multidimensional vector. The learning process consists of updating the hidden layers weights based on a loss function evaluated on the output. Having multiple hidden layers stacked upon each-other is oftenly referred to as DL.

CNNs are analogous to traditional ANNs in that they are composed of neurons that are optimized through learning. The key difference between CNNs and ANNs are that neurons are connected only locally and the remaining weights are shared reproducing the convolution operation. This results into a reduction in the number

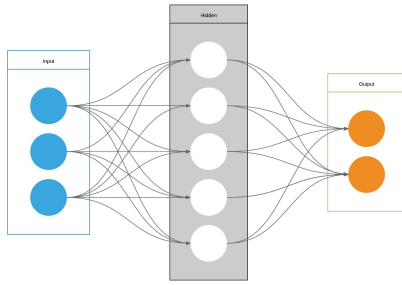

Figure A.1. Artificial Neural Network representation.

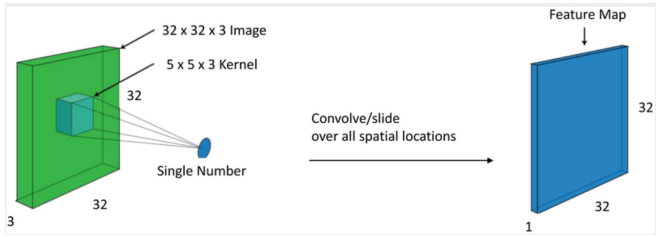

Figure A.2. Convolution operation in a sliding window fashion to obtain an activation/feature map.

of parameters, and some translational invariance. By using multiple convolutional filters in a hierarchical fashion, CNNs encode image-specific features, making this type of architecture better suited for image-analysis tasks. Reducing the number of parameters is beneficial for two reasons: (i) usually we do not have unlimited computational power and time to train these models, and (ii) models with a large number of parameters are prompter to overfitting.

A.2 Convolutional Neural Networks

CNNs are comprised of three main elements: convolutional layers, pooling layers, and fully-connected layers. A CNN architecture is a stack of these layers. More advanced elements such as dropout layers

or batch normalization layers. Investigating different type of layers is an active area of research.

Convolutional layer

Convolutional layers are the key component of CNNs. The layers' parameters define the weights of learnable convolutional kernels.

Each layer l is parameterized by four variables:

- k_l: kernel size (positive integer)
- s_l: stride size (positive integer)
- p_l: padding applied to the sides of the input feature map (non-negative integer)

These kernels are usually small, and are applied in a sliding window fashion across the input image. This layer convolves each filter with a window of the input image, producing as an output a 2D *activation feature map*. The network will learn kernels that produce a high activation at given spatial positions of the input. Unlike ANN, neurons in the CNN are only connected to a region of the input. Parameter sharing works on the assumption that features found relevant on one region, should also be relevant on other image regions. The dimensionality of this region is commonly referred to as the *receptive field* of the neuron. At every convolutional layer, several kernels are learned and their corresponding activation maps are stacked along the depth dimension.

Pooling layer

Pooling layers aim to gradually reduce the dimensionality of the representation. Thereby, also reducing the number of parameters and the complexity of the network. The two most common pooling operations are the "max" f_{max} or "average" f_{avg} functions. These two are simple functions. For each kernel size k_l in the input feature,

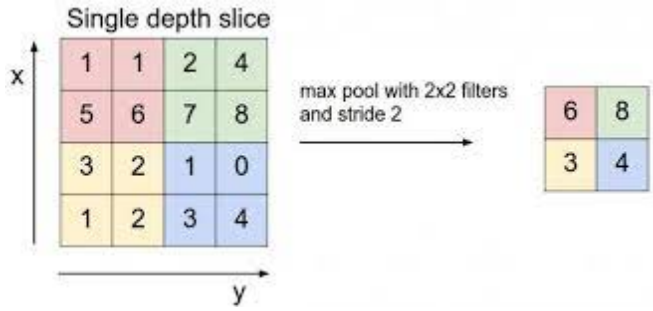

Figure A.3. Example of a max pool operation with kernel size $k_l = 2$ and stride $s_l = 2$.

f_{max} considers as output the maximum element within the region, while f_{avg} computes the average. Although pooling layers perform a different operation than convolutional layers, they are characterized by the same parameters: kernel size k_l, stride size s_l, and padding p_l.

Fully-connected layer

A fully-connected layer is composed of neurons connected to every other neuron from the previous adjacent layer. This is analogous to the way neurons are arranged in traditional forms of ANN (Figure A.1).

Dropout layer

Dropout is a regularization technique for reducing overfitting in CNNs. The term dropout refers to randomly "dropping out" (or omitting) neurons during the training process of a neural network [SHK+14, WFGCB13]. It is sometimes also referred as *dilution*, which stands for thinning of the weights.

Batch normalization layer

Batch normalization layers [IS15] were initially introduced to mitigate *internal convariate shift*. The phenomenon of internal covariate shift arises due to the randomness in the parameter initialization and the randomness in the input data, which affect the layer's output distribution. During training, the mean and variance distribution of the inputs to each layer change accordingly to the input batch. However, recently Santurkar *et al.* [STIM18] argued that batch normalization rather smooths the objective function, resulting into a more predictive and stable behaviour of the gradients, allowing for faster training.

Batch normalization layers present other benefits. One of them is that the network can be trained with a larger learning rate without the problem of vanishing gradients. Moreover, the regularizing effect improves the network generalization properties, and it can reduce overfitting. After batch normalization layer, many other in-layer normalization methods have been introduced, such as instance normalization [UVL16], layer normalization [BKH16] and group normalization [WH18].

Bibliography

[AAHR13] Mahmoud Al-Ayyoub, Ismail Hmeidi, and Haya Rababah. Detecting hand bone fractures in x-ray images. *JMPT*, 4(3):155–168, 2013.

[ABA+16] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, May 2016.

[AdTBT20] Mathieu Andreux, Jean Ogier du Terrail, Constance Beguier, and Eric W. Tramel. Siloed federated learning for multi-centric histopathology datasets. In Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu, editors, *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 129–139, Cham, 2020. Springer International Publishing.

[AEBS+20] Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. A curriculum learning based approach to captioning ultrasound images. In Yipeng Hu, Roxane Licandro, J. Alison Noble, Jana Hutter, Stephen Aylward, Andrew Melbourne, Esra Abaci Turk, and Jordina Torrents Barrena, editors, *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 75–84, Cham, 2020. Springer International Publishing.

[AKF+20] Euijoon Ahn, Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. *IEEE Transactions on Medical Imaging*, 39(7):2385–2394, 2020.

[Ann03] George J Annas. Hipaa regulations—a new era of medical-record privacy?, 2003.

[BÇ16] Fatih Bayram and Murat Çakıroğlu. Diffract: Diaphyseal femur fracture classifier system. *Biocybernetics and Biomedical Engineering*, 36(1):157–171, 2016.

[BDE+15] M. Bhandari, P. J. Devereaux, T. A. Einhorn, L. Thabane, E. H. Schemitsch, K. J. Koval, F. Frihagen, R. W. Poolman, K. Tetsworth, E. Guerra-Farfan, K. Madden, S. Sprague, G. Guyatt, and HEALTH Investigators. Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial. *BMJ Open*, 5(2):e006263–e006263, feb 2015.

[BIK+17] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel,

Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[BLC14] Umberto Giovanni De Bellis, Claudio Legnani, and Giorgio Maria Calori. Acute total hip replacement for acetabular fractures: A systematic review of the literature. *Injury*, 45(2):356–361, feb 2014.

[BLCW09] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM.

[BLZ+18] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[BRPC+20] Ilaria Bonavita, Xavier Rafael-Palou, Mario Ceresa,

Gemma Piella, Vicent Ribas, and Miguel A. González Ballester. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. *Computer Methods and Programs in Biomedicine*, 185:105172, 2020.

[BSD+17] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[BWRA21] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Shadi Albarqouni. Feddis: Disentangled federated learning for unsupervised brain pathology segmentation. *arXiv preprint arXiv:2103.03705*, 2021.

[BWY92] Richard E Bird, Terry W Wallace, and Bonnie C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613–617, 1992.

[BZOR+19] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1), April 2019.

[CAL+17] M. Jorge Cardoso, Tal Arbel, Su-Lin Lee, Veronika Cheplygina, Simone Balocco, Diana Mateus, Guillaume Zahnd, Lena Maier-Hein, Stefanie Demirci, Eric Granger, Luc Duong, Marc-André Carbonneau, Shadi Albarqouni, and Gustavo Carneiro, editors. *Intravascular Imaging and Computer Assisted Stenting, and*

Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 6th Joint International Workshops, CVII-STENT and Second International Workshop, LABELS, 2017. Held in Conjunction with MICCAI 2017.

[CHL+19] Chi-Tung Cheng, Tsung-Ying Ho, Tao-Yi Lee, Chih-Chen Chang, Ching-Cheng Chou, Chih-Chi Chen, I-Fang Chung, and Chien-Hung Liao. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *European Radiology*, 29(10):5469–5477, Oct 2019.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[CP18] Veronika Cheplygina and Josien P. W. Pluim. Crowd disagreement about medical images is informative. In Danail Stoyanov, Zeike Taylor, Simone Balocco, Raphael Sznitman, Anne Martel, Lena Maier-Hein, Luc Duong, Guillaume Zahnd, Stefanie Demirci, Shadi Albarqouni, Su-Lin Lee, Stefano Moriconi, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, Eric Granger, and Pierre Jannin, editors, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label*

Synthesis, pages 105–111, Cham, 2018. Springer International Publishing.

[CPC+17] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Just dial: Domain alignment layers for unsupervised domain adaptation. In *International Conference on Image Analysis and Processing*, pages 357–369. Springer, 2017.

[CYS+19] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[D'O96] Carl J D'Orsi. The american college of radiology mammography lexicon: an initial attempt to standardize terminology. *AJR. American journal of roentgenology*, 166(4):779–780, 1996.

[Doi07] 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, June 2007.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[DTC+02] Stephen W Duffy, Laszlo Tabár, Hsiu-Hsi Chen, Marit Holmqvist, Ming-Fang Yen, Shahim Abdsalah, Birgitta Epstein, Ewa Frodis, Eva Ljungberg, Christina Hedborg-Melander, et al. The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties: a collaborative evaluation. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 95(3):458–469, 2002.

[dVWdJ+16] Bob D. de Vos, Jelmer M. Wolterink, Pim A. de Jong, Max A. Viergever, and Ivana Išgum. 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In Martin A. Styner and Elsa D. Angelini, editors, *Medical Imaging 2016: Image Processing*, volume 9784, pages 517 – 523. International Society for Optics and Photonics, SPIE, 2016.

[EHBG07] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 127–136, New York, NY, USA, 2007. ACM.

[EKM+18] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems*, pages 842–852, 2018.

[EKN+17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin

Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, jan 2017.

[Elm93] Jeffrey L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, jul 1993.

[FPW+00] Patrick L Fitzgibbons, David L Page, Donald Weaver, Ann D Thor, D Craig Allred, Gary M Clark, Stephen G Ruby, Frances O'Malley, Jean F Simpson, James L Connolly, et al. Prognostic factors in breast cancer: College of american pathologists consensus statement 1999. *Archives of pathology & laboratory medicine*, 124(7):966–978, 2000.

[FSA99] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[FSST97] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168, Aug 1997.

[FWD+18] E. R. Flikweert, K. W. Wendt, R. L. Diercks, G. J. Izaks, D. Landsheer, M. Stevens, and I. H. F. Reininga. Complications after hip fracture surgery: are they preventable? *European journal of trauma and emergency surgery : official publication of the European Trauma Society*, 44(4):573–580, Aug 2018. 28795198[pmid].

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Int. Conf. on Artificial Intelligence and Statistics*, volume 9, pages 249–256. PMLR, 13–15 May 2010.

[GCG16] Dionysios Giannoulis, Giorgio M. Calori, and Peter V. Giannoudis. Thirty-day mortality after hip fractures: has anything changed? *European Journal of Orthopaedic Surgery & Traumatology*, 26(4):365–370, mar 2016.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[GGG+19] Florin C. Ghesu, Bogdan Georgescu, Eli Gibson, Sebastian Guendel, Mannudeep K. Kalra, Ramandeep Singh, Subba R. Digumarthy, Sasa Grbic, and Dorin Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 676–684, Cham, 2019. Springer International Publishing.

[GHJK20] Felix Grimberg, Mary-Anne Hartley, Martin Jaggi, and Sai Praneeth Karimireddy. Weight erosion: An update aggregation scheme for personalized collaborative machine learning. In Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu, editors, *Domain Adaptation and Representation Transfer, and*

Distributed and Collaborative Learning, pages 160–169, Cham, 2020. Springer International Publishing.

[Gir15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[GJR14] Konstantin V Grigoryan, Houman Javedan, and James L Rudolph. Ortho-geriatric care models and outcomes in hip fracture patients: a systematic review and meta-analysis. *Journal of orthopaedic trauma*, 28(3):e49, 2014.

[GL15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[GLS+18] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113 – 122, 2018.

[GPAM+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[GUA+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[Hal12] Kevin A Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.

[HG16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

[HGCB16] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.

[HHIL+17] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.

[HK16] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for fully weakly supervised object localization. *CoRR*, abs/1602.01625, 2016.

[HLCLT16] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016.

[HLL+20] W. Huang, M. Luo, X. Liu, P. Zhang, and H. Ding. Arterial spin labeling image synthesis from structural mri using improved capsule-based networks. *IEEE Access*, 8:181137–181153, 2020.

[HM15] Eric Horvitz and Deirdre Mulligan. Data, privacy, and the greater good. *Science*, 349(6245):253–255, July 2015.

[HMZ18] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[HSF18] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018.

[HSQ+19] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[HW19] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *arXiv preprint arXiv:1904.03626*, 2019.

[HZKH20] Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16797–16807. Curran Associates, Inc., 2020.

[HZRS16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[HZRS16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[JDC+05] Wen-Jie Jin, Li-Yang Dai, Yi-Min Cui, Qing Zhou, Lei-Sheng Jiang, and Hua Lu. Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. *Injury*, 36(7):858–861, jul 2005.

[JGG+17] Andrew Jesson, Nicolas Guizard, Sina Hamidi Ghalehjegh, Damien Goblot, Florian Soudan, and Nicolas Chapados. CASED: Curriculum adaptive sampling for extreme data imbalance. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 639–646, Cham, 2017. Springer International Publishing.

[JKA+19] Amelia Jiménez-Sánchez, Anees Kazi, Shadi Albarqouni, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Diana Mateus, and Sonja Kirchhoff. Towards an interactive and interpretable CAD system to sup-

port proximal femur fracture classification. *CoRR*, abs/1902.01338v1, 2019.

[JMY+14] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2078–2086. Curran Associates, Inc., 2014.

[Jøs18] Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 1st edition, 2018.

[JSAM18] Amelia Jiménez-Sánchez, Shadi Albarqouni, and Diana Mateus. Capsule networks against medical imaging data challenges. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 150–160, Cham, 2018. Springer International Publishing.

[JSKA+20] Amelia Jiménez-Sánchez, Anees Kazi, Shadi Albarqouni, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Sonja Kirchhoff, and Diana Mateus. Precise proximal femur fracture classification for interactive training and surgical planning. *International Journal of Computer Assisted Radiology and Surgery*, 15(5):847–857, April 2020.

[JSMK+19] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline

Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 694–702, Cham, 2019. Springer International Publishing.

[JSMK+21] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. Curriculum learning for improved femur fracture classification: scheduling data with prior knowledge and uncertainty. *Medical Image Analysis (submitted)*, 2021.

[JSTB+21] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A González Ballester, Diana Mateus, and Gemma Piella. Memory-aware curriculum federated learning for breast cancer classification. *IEEE Transactions on Medical Imaging (submitted)*, 2021.

[JZL+18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[KAS+17a] Anees Kazi, Shadi Albarqouni, Amelia Jimenez Sanchez, Sonja Kirchhoff, Peter Biberthaler, Nassir Navab, and Diana Mateus. Automatic classification of proximal femur fractures based on attention models. In Qian Wang, Yinghuan Shi, Heung-Il Suk, and Kenji Suzuki, editors, *Machine Learning in Medical Imaging*, pages 70–78, Cham, 2017. Springer International Publishing.

[KAS+17b] Anees Kazi, Shadi Albarqouni, Amelia Jimenez Sanchez, Sonja Kirchhoff, Peter Biberthaler, Nassir Navab, and Diana Mateus. Automatic classification of proximal femur fractures based on attention models. In Qian Wang, Yinghuan Shi, Heung-Il Suk, and Kenji Suzuki, editors, *Machine Learning in Medical Imaging*, pages 70–78, Cham, 2017. Springer International Publishing.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[KCG+00] Karla Kerlikowske, Patricia A Carney, Berta Geller, Margaret T Mandelson, Stephen H Taplin, Kathy Malvin, Virginia Ernster, Nicole Urban, Gary Cutter, Robert Rosenberg, et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. *Annals of internal medicine*, 133(11):855–863, 2000.

[KCM19] Gene Kitamura, Chul Y. Chung, and Barry E. Moore. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. *Journal of Digital Imaging*, 32(4):672–677, Aug 2019.

[KDGBA19] Hoel Kervadec, Jose Dolz, Éric Granger, and Ismail Ben Ayed. Curriculum semi-supervised segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 568–576, Cham, 2019. Springer International Publishing.

[KKK+07] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietila. the diaretdb1 diabetic retinopathy database and evaluation protocol. In *Proc. BMVC*, pages 15.1–15.10, 2007. doi:10.5244/C.21.15.

[KkKV+] V. Kalesnykiene, J. k. Kamarainen, R. Voutilainen, J. Pietilä, H. Kälviäinen, and H. Uusitalo. Diaretdb1 diabetic retinopathy database and evaluation protocol.

[KMA+18] James F. Kellam, Eric G. Meinberg, Julie Agel, Matthew D. Karam, and Craig S. Roberts. Introduction. *Journal of Orthopaedic Trauma*, 32:S1–S10, jan 2018.

[KPK10] M. P. Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1189–1197. Curran Associates, Inc., 2010.

[KSH12a] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[KSH12b] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[KSTH19] Adam R Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked capsule autoencoders. *arXiv preprint arXiv:1906.06818*, 2019.

[LA17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[LB18a] Rodney LaLonde and Ulas Bagci. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018.

[LB18b] Rodney LaLonde and Ulas Bagci. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[LC10] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[LCWJ15a] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[LCWJ15b] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[LGD+20] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020.

[LGG+17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[LGH+17] 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(1), December 2017.

[LJZ+21] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

[LKB+17a] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm 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.

[LKB+17b] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso 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. *CoRR*, abs/1702.05747, 2017.

[LLT19] Yingting Li, Lu Liu, and Robby T. Tan. Certainty-driven consistency loss for semi-supervised learning. *CoRR*, abs/1901.05657, 2019.

[LMX+19] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan

Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141. Springer, 2019.

[LSZ+18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

[LTB20] Rodney LaLonde, Drew Torigian, and Ulas Bagci. Encoding visual attributes in capsules for explainable medical diagnoses. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 294–304. Springer, 2020.

[LTY13] M. Lin, K. Tang, and X. Yao. Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):647–660, April 2013.

[LWD+13] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *2013 IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

[MA18] Melike Nur Mermer and Mehmet Fatih Amasyali. Training with growing sets: A simple alternative to curriculum learning and self paced learning, 2018.

[MAD+12] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. INbreast. *Academic Radiology*, 19(2):236–248, February 2012.

[MARC20] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. *arXiv preprint arXiv:2007.12256*, 2020.

[MB12] Ravi Mittal and Sumit Banerjee. Proximal femoral fractures: Principles of management and review of literature. *Journal of Clinical Orthopaedics and Trauma*, 3(1):15–23, June 2012.

[MBN+18] Gabriel Maicas, Andrew P Bradley, Jacinto C Nascimento, Ian Reid, and Gustavo Carneiro. Training medical image analysis systems like radiologists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–554. Springer, 2018.

[MBTR18] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018.

[Mer14] Cecilia L. Mercado. BI-RADS update. *Radiologic Clinics of North America*, 52(3):481–487, May 2014.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[MMR+17a] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[MMR+17b] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[MOCS17] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *CoRR*, abs/1707.00183, 2017.

[MR18] Payman Mohassel and Peter Rindal. Aby3: A mixed protocol framework for machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 35–52, 2018.

[MSMK10] Subhasis Misra, Naveenraj L Solomon, Frederick L Moffat, and Leonidas G Koniaris. Screening criteria for breast cancer. *Advances in surgery*, 44(1):87–100, 2010.

[MYCVN20] Aryan Mobiny, Pengyu Yuan, Pietro Antonio Cicalese, and Hien Van Nguyen. Decaps: Detail-oriented capsule networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 148–158. Springer, 2020.

[MZJ17] Deyu Meng, Qian Zhao, and Lu Jiang. A theoretical understanding of self-paced learning. *Information Sciences*, 414:319 – 328, 2017.

[NSW16] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016.

[O+20] World Health Organization et al. Diabetic retinopathy screening: a short guide. 2020.

[OFM+17] Jakub Olczak, Niklas Fahlberg, Atsuto Maki, Ali Sharif Razavian, Anthony Jilert, André Stark, Olof Sköldenberg, and Max Gordon. Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms—are they on par with humans for diagnosing fractures? *Acta orthopaedica*, 88(6):581–586, 2017.

[ON15] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[PCL+19] Beomhee Park, Yongwon Cho, Gaeun Lee, Sang Min Lee, Young-Hoon Cho, Eun Sol Lee, Kyung Hee Lee, Joon Beom Seo, and Namkug Kim. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-pa x-ray screening for pulmonary abnormalities. *Scientific reports*, 9(1):1–9, 2019.

[PHZS19] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019.

[PKM19] Daniel Peterson, Pallika Kanani, and Virendra J Marathe. Private federated learning with domain adaptation. *arXiv preprint arXiv:1912.06733*, 2019.

[PKP+99] Jari Parkkari, Pekka Kannus, Mika Palvanen, A Natri, J Vainio, H Aho, I Vuori, and M Järvinen. Majority of hip fractures occur as a result of a fall and impact on the greater trochanter of the femur: a prospective controlled hip fracture study with 206 consecutive patients. *Calcified tissue international*, 65(3):183–187, 1999.

[PRW+17] Daniel Pincus, Bheeshma Ravi, David Wasserstein, Anjie Huang, J. Michael Paterson, Avery B. Nathens,

Hans J. Kreder, Richard J. Jenkinson, and Walter P. Wodchis. Association Between Wait Time and 30-Day Mortality in Adults Undergoing Hip Fracture Surgery. *JAMA*, 318(20):1994–2003, 11 2017.

[QCSLS09] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[RCS+17] Abhijit Guha Roy, Sailesh Conjeti, Debdoot Sheet, Amin Katouzian, Nassir Navab, and Christian Wachinger. Error corrective boosting for learning fully convolutional networks with limited data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 231–239. Springer, 2017.

[RCS+20] Holger R Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C Bizzo, et al. Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 181–191. Springer, 2020.

[RHdM19] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), July 2019.

[RHL+20a] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):1–7, 2020.

[RHL+20b] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1), September 2020.

[RHS+18] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 201–209, Cham, 2018. Springer International Publishing.

[RHU+18] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports*, 8(1), March 2018.

[RLA+15] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[RWY+16] Holger R Roth, Yinong Wang, Jianhua Yao, Le Lu, Joseph E Burns, and Ronald M Summers. Deep convolutional networks for automated detection of posterior-element fractures on spine ct. In *Medical Imaging 2016:*

Computer-Aided Diagnosis, volume 9785, page 97850P. International Society for Optics and Photonics, 2016.

[RYY+15] Devon J. Ryan, Hiroyuki Yoshihara, Daisuke Yoneoka, Kenneth A. Egol, and Joseph D. Zuckerman. Delay in hip fracture surgery: An analysis of patient-specific and hospital-specific risk factors. *Journal of Orthopaedic Trauma*, 29(8):343–348, aug 2015.

[RYY+20] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020.

[SBWM08] Mehul P Sampat, Alan C Bovik, Gary J Whitman, and Mia K Markey. A model-based framework for the detection of spiculated masses on mammography a. *Medical physics*, 35(5):2110–2123, 2008.

[Sch03] Robert E. Schapire. *The Boosting Approach to Machine Learning: An Overview*, pages 149–171. Springer New York, New York, NY, 2003.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[SCLW19] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4951–4958, 2019.

[SCN+18] Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran,

Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, et al. Medal: Accurate and robust deep active learning for medical image analysis. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 481–488. IEEE, 2018.

[SDBR14] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[SFH17] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3856–3866. Curran Associates, Inc., 2017.

[SFS+21] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, February 2021.

[SG18] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

[Sha48] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[SHK+14] 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.

[SKK18] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[SMR+19] 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), August 2019.

[SP98] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, Jan 1998.

[SRE+18] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104. Springer, 2018.

[SRG+16a] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[SRG+16b] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[SSCY19] Hai Su, Xiaoshuang Shi, Jinzheng Cai, and Lin Yang. Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 559–567, Cham, 2019. Springer International Publishing.

[SSW+15] Scott E. Sheehan, Jeffrey Y. Shyu, Michael J. Weaver, Aaron D. Sodickson, and Bharti Khurana. Proximal femoral fractures: What the orthopedic surgeon wants to know. *RadioGraphics*, 35(5):1563–1584, September 2015.

[SSY+19] Hong Shang, Zhongqian Sun, Wei Yang, Xinghui Fu, Han Zheng, Jia Chang, and Junzhou Huang. Leveraging other datasets for medical imaging classification: Evaluation of transfer, multi-task and semi-supervised learning. In Dinggang Shen, Tianming Liu, Terry M.

Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 431–439, Cham, 2019. Springer International Publishing.

[STD19] Shreyas Saxena, Oncel Tuzel, and D. DeCoste. Data parameters: A new family of parameters for learning a differentiable curriculum. In *NeurIPS*, 2019.

[STIM18] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *arXiv preprint arXiv:1805.11604*, 2018.

[STY+20] S. Sabour, Andrea Tagliasacchi, S. Yazdani, Geoffrey E. Hinton, and David J. Fleet. Unsupervised part representation by flow capsules. *ArXiv*, abs/2011.13920, 2020.

[Sum10] Ronald M. Summers. Improving the accuracy of CTC interpretation: Computer-aided detection. *Gastrointestinal Endoscopy Clinics of North America*, 20(2):245–257, April 2010.

[SWH84] MARC F Swiontkowski, RA Winquist, and Jr ST Hansen. Fractures of the femoral neck in patients between the ages of twelve and forty-nine years. *The Journal of bone and joint surgery. American volume*, 66(6):837–846, 1984.

[SWS17] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. PMID: 28301734.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[TIM+20] Agnieszka Tomczak, Slobodan Ilic, Gaby Marquardt, Thomas Engel, Frank Forster, Nassir Navab, and Shadi Albarqouni. Multi-task multi-domain learning for digital staining and classification of leukocytes. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020.

[TNA19] Agnieszka Tomczack, Nassir Navab, and Shadi Albarqouni. Learn to estimate labels uncertainty for quality assurance. *ArXiv*, abs/1909.08058, 2019.

[TSG+16] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[tup] MICCAI Grand Challenge Tumor Proliferation Assessment Challenge (TUPAC16). `http://tupac.tue-image.nl/`. Accessed: 2018-01-18.

[TVM+20] Leonardo Tanzi, Enrico Vezzetti, Rodrigo Moreno, Alessandro Aprato, Andrea Audisio, and Alessandro Massè. Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach. *European Journal of Radiology*, 133:109373, 2020.

[TWH+18] Yuxing Tang, Xiaosong Wang, Adam P. Harrison, Le Lu, Jing Xiao, and Ronald M. Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *Machine Learning in Medical Imaging*, pages 249–258, Cham, 2018. Springer International Publishing.

[UTG+19] Takaaki Urakawa, Yuki Tanaka, Shinichi Goto, Hitoshi Matsuzawa, Kei Watanabe, and Naoto Endo. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiology*, 48(2):239–244, Feb 2019.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[vDvdWB04] Paul J van Diest, Elsken van der Wall, and Jan PA Baak. Prognostic value of proliferation in invasive breast cancer: a review. *Journal of clinical pathology*, 57(7):675–681, 2004.

[vERMR10] D. van Embden, S.J. Rhemrev, S.A.G. Meylaerts, and G.R. Roukema. The comparison of two classifications for trochanteric femur fractures: The AO/ASIF classification and the jensen classification. *Injury*, 41(4):377–381, apr 2010.

[vGSJC15] Bram van Ginneken, Arnaud A. A. Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th In-*

ternational Symposium on Biomedical Imaging (ISBI), pages 286–289, 2015.

[vGvGH+16] Mark J. J. P. van Grinsven, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, and Clara I. Sanchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, may 2016.

[VPS+16] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, aug 2016.

[VV17a] C Nader Vasconcelos and B Nader Vasconcelos. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. *CoRR, abs/1702.07025*, 1, 2017.

[VV17b] Cristina Nader Vasconcelos and Bárbara Nader Vasconcelos. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. *CoRR*, abs/1702.07025, 2017.

[VVDJ+16] Mitko Veta, Paul J Van Diest, Mehdi Jiwa, Shaimaa Al-Janabi, and Josien PW Pluim. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS one*, 11(8):e0161286, 2016.

[WCA18] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and

experiments with deep networks. *arXiv preprint arXiv:1802.03796*, 2018.

[WDN21] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations*, 2021.

[WDW+12] Jie Wu, Pavani Davuluri, Kevin R Ward, Charles Cockrell, Rosalyn Hobson, and Kayvan Najarian. Fracture detection in traumatic pelvic ct images. *Journal of Biomedical Imaging*, 2012:1, 2012.

[WFGCB13] David Warde-Farley, Ian J Goodfellow, Aaron Courville, and Yoshua Bengio. An empirical analysis of dropout in piecewise linear networks. *arXiv preprint arXiv:1312.6197*, 2013.

[WGY+19] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[WH18] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[WLC+19a] Yirui Wang, Le Lu, Chi-Tung Cheng, Dakai Jin, Adam P Harrison, Jing Xiao, Chien-Hung Liao, and Shun Miao. Weakly supervised universal fracture detection in pelvic x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 459–467. Springer, 2019.

[WLC+19b] Yirui Wang, Le Lu, Chi-Tung Cheng, Dakai Jin, Adam P. Harrison, Jing Xiao, Chien-Hung Liao, and Shun Miao. Weakly supervised universal fracture detection in pelvic x-rays. In Dinggang Shen, Tianming

Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 459–467, Cham, 2019. Springer International Publishing.

[WLD20] Zhao Wang, Quande Liu, and Qi Dou. Contrastive cross-site learning with redesigned net for covid-19 ct classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2806–2813, 2020.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[WPP+19] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2019.

[WRL+18] J. Wu, S. Ruan, C. Lian, S. Mutic, M. A. Anastasio, and H. Li. Active learning with noise modeling for medical image annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 298–301, April 2018.

[WS19] Tien Y. Wong and Charumathi Sabanayagam. The war on diabetic retinopathy: Where are we now? *Asia-Pacific Journal of Ophthalmology*, 8(6):448–456, November 2019.

[WSMM18] Ken C.L. Wong, Tanveer Syeda-Mahmood, and Mehdi Moradi. Building medical image classifiers with very limited data using segmentation networks. *Medical Image Analysis*, 49:105 – 116, 2018.

[XDS+19] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1280–1283. IEEE, 2019.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[YBLS20] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. *arXiv preprint arXiv:2007.01261*, 2020.

[YFNA20] Yousef Yeganeh, Azade Farshad, Nassir Navab, and Shadi Albarqouni. Inverse distance aggregation for federated learning with non-iid data. In Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu, editors, *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 150–159, Cham, 2020. Springer International Publishing.

[YWL+19] J. Yang, X. Wu, J. Liang, X. Sun, M. Cheng, P. L. Rosin, and L. Wang. Self-paced balance learning for clinical skin disease recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2019.

[ZB18] Tianyi Zhou and Jeff Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *International Conference on Learning Representations*, 2018.

[ZCCL20] Rongchang Zhao, X. Chen, Zailiang Chen, and Shuo Li. Egdcl: An adaptive curriculum learning framework for unbiased glaucoma diagnosis. 2020.

[ZCDLP17] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[ZLZ+17a] Jianlong Zhou, Zelin Li, Weiming Zhi, Bin Liang, Daniel Moses, and Laughlin Dawes. Using convolutional neural networks and transfer learning for bone age classification. *Int. Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, 2017.

[ZLZ+17b] Jianlong Zhou, Zelin Li, Weiming Zhi, Bin Liang, Daniel Moses, and Laughlin Dawes. Using convolutional neural networks and transfer learning for bone age classification. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2017.

[ZNH+18] Lingchen Zhao, Lihao Ni, Shengshan Hu, Yaniiao Chen, Pan Zhou, Fu Xiao, and Libing Wu. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 2087–2095. IEEE, 2018.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using

cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[Zuc96] Joseph D Zuckerman. Hip fracture. *New England journal of medicine*, 334(23):1519–1525, 1996.

[ZWW+20] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.

[ZZ09] Ethan Zhang and Yi Zhang. *Eleven Point Precision-recall Curve*, pages 981–982. Springer US, Boston, MA, 2009.