Meta-learning for Clinical and Imaging Data Fusion for Improved Deep Learning Inference

Kirill Vasilevski
Department of Electrical and Computer Engineering
McGill University, Montreal
February, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

Deep learning methods such as convolutional neural networks (CNN) have achieved state-of-the-art success in a variety of medical imaging applications such as pathology segmentation, diagnosis, and prediction of prognosis using information from complex imaging data (e.g. CT, MRI). However, deep learning models can still make mistakes when performing predictions based on images alone, leading to potentially devastating consequences if embedded into real clinical workflows. In medical practice, imaging findings are often interpreted in combination with clinical context provided by non-imaging data, resulting in more informative decision making and improved diagnostic accuracy. How to best leverage additional clinical and other information in order to improve image-based deep learning models remains an open research problem. In this thesis, we propose a meta-learning deep learning method for interpretation of medical images that guides a CNN model to learn imaging features that are informed by clinical context, allowing for improved classification performance over imaging-only methods in the medical imaging domain. A CNN model is pretrained using a meta-learning scheme in order to learn shared imaging features that are predictive of several relevant clinical markers and are informative of the clinical context, following which the model can be adapted to a related desired clinical task (e.g. diagnosis). To evaluate our method, we leverage Multiple Sclerosis (MS) and Alzheimer's Disease (AD) datasets containing 3D MRI and tabular clinical data to perform three clinical tasks: detection of brain lesion presence and prediction of future lesional activity in MS patients, as well as regression of neurological assessment scores in AD patients. We compare our method against existing deep learning methods

that use imaging-only data, clinical-only data, and a combination of the two to analyze importance of individual data modalities along with various ways of combining them. Furthermore, we perform an exploratory ablation study on the selection of supporting clinical data, role of dataset size, and their impacts on model performance. Based on the selected metrics, the proposed approach performed better compared to imaging-only methods across all tasks, as well as achieved better or comparable classification and regression performance compared to other data fusion methods on select tasks. Lastly, our approach showed robust performance across varying dataset sizes compared to other baselines, making it more suitable for real-world applications where there is often a lack of available training data. The findings outlined in this thesis can be of use to other researchers by demonstrating how meta-learning methods can help train more informative medical imaging models.

Abrégé

Les méthodes d'apprentissage profond telles que les réseaux neuronaux convolutifs (CNN) ont connu un succès de pointe dans une variété d'applications d'imagerie médicale telles que la segmentation des pathologies, le diagnostic et la prédiction du pronostic à partir d'informations provenant de données d'imagerie complexes (par exemple, CT, IRM). Cependant, les prédictions d'apprentissage profond peuvent encore faire des erreurs lorsqu'elles effectuent des prédictions basées uniquement sur des images, ce qui entraîne des conséquences potentiellement dévastatrices si elles sont intégrées dans les cliniques réels. Dans la pratique médicale, les résultats d'imagerie sont souvent interprétés en combinaison avec le contexte clinique fourni par des données non liées à l'imagerie, ce qui permet de prendre des décisions plus informatives et d'améliorer la précision du diagnostic. La façon d'exploiter au mieux les informations cliniques supplémentaires et autres afin d'améliorer les modèles d'apprentissage profond basés sur l'image reste un problème de recherche ouvert. Dans cette thèse, nous proposons une méthode de méta-apprentissage profond pour l'interprétation des images médicales qui guide le modèle CNN pour apprendre les caractéristiques d'imagerie qui sont informées par le contexte clinique, ce qui permet d'améliorer les performances de classification par rapport aux méthodes basées uniquement sur l'imagerie dans le domaine de l'imagerie médicale. Un modèle CNN est pré-entraîné à l'aide d'un schéma de méta-apprentissage afin d'apprendre des caractéristiques d'imagerie partagées qui sont prédictives de plusieurs marqueurs cliniques pertinents et qui sont informatives du contexte clinique, après quoi le modèle peut être adapté à une tâche clinique connexe souhaitée (par exemple, le diagnostic). Pour évaluer notre

méthode, nous exploitons des ensembles de données sur la sclérose en plaques (SEP) et la maladie d'Alzheimer (MA) contenant des données IRM 3D et des données cliniques tabulaires pour réaliser trois tâches cliniques: la détection de la présence de lésions cérébrales et la prédiction de l'activité lésionnelle future chez les patients atteints de SEP, ainsi que la régression des scores d'évaluation neurologique chez les patients atteints de MA. Nous comparons notre méthode aux méthodes d'apprentissage profond existantes qui utilisent des données d'imagerie uniquement, des données cliniques uniquement et une combinaison des deux, afin d'analyser l'importance de chaque modalité de données ainsi que les différentes façons de les combiner. En outre, nous réalisons une étude exploratoire sur la sélection des données cliniques de soutien, le rôle de la taille de l'ensemble de données, et leurs impacts sur la performance du modèle. Sur la base de certaines mesures, l'approche proposée a donné de meilleurs résultats que les méthodes basées uniquement sur l'imagerie pour toutes les tâches, et a obtenu des performances de classification et de régression supérieures ou comparables à celles d'autres méthodes de fusion de données pour certaines tâches. Enfin, notre approche a montré des performances robustes sur des ensembles de données de taille variable, par rapport aux autres méthodes de référence, ce qui la rend plus appropriée pour les applications du monde réel où il y a souvent un manque de données d'entraînement disponibles. Les résultats présentés dans cette thèse peuvent être utiles à d'autres chercheurs en démontrant comment les méthodes de méta-apprentissage peuvent aider à former des modèles d'imagerie médicale plus informatifs.

Acknowledgements

I would like to first thank my family, Svetlana, Andrei, and Anna for all the continued love and support throughout the years. I would also like to sincerely thank my partner, Victoria, for always being there to support and encourage me, as well as braving countless train trips to Montreal for the past two years. I also thank all my friends for keeping me emotionally sane throughout this endeavour.

Furthermore, sincerest thank you to my supervisor, Dr. Tal Arbel, for all the guidance, insights, and support throughout my Master's studies. I also extend my deepest gratitude to associate member, Dr. Behrooz Mahasseni, for the countless discussions, collaborations, and mentoring. This thesis would not have been possible without you and all your help.

To my friends and lab members, Eric Zimmermann, Justin Szeto, and Jillian Cardinell, thank you for all your invaluable support and camaraderie. From braving tough courses, to exchanging research ideas, all-night hackathons, code reviews, and encouraging me through challenging times. Without you, I sincerely would not have been able to complete this program, and attribute a lot of my personal and academic growth you. Thank you.

Lastly, I am grateful to the companies who generously provided the Multiple Sclerosis clinical trial data that made this research possible: Biogen, Hoffman-La Roche, and Teva. This investigation was supported (in part) by an award from the International Progressive Multiple Sclerosis Alliance (award reference number PA-1412-02420). Additionally, I would like to thank Louis Collins and Mahsa Dadar for preprocessing the MRI data, Zografos Caramanos, Alfredo Morales Pinzon, Charles Guttmann and István Morocz for

collating the clinical data, and Sridar Narayanan. Funding was also provided by Natural Sciences and Engineering Research Council of Canada (Arbel) and the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs (Arbel). I am also grateful to the Alzheimer's Disease Neuroimaging Initiative (ADNI; adni.loni.usc.edu) for providing Alzheimer's disease data for this research. Supplementary computational resources and technical support were provided by Calcul Québec, WestGrid, and Compute Canada.

Table of Contents

List of Figures

List of Tables

List of Acronyms

DL	Deep Learning
MS	Multiple Sclerosis
AD	Alzheimer's Disease
EDSS	Expanded Disability Status Scale
RRMS	Relapsing Remitting Multiple Sclerosis
PPMS	Primary Progressive Multiple Sclerosis
SPMS	Secondary Progressive Multiple Sclerosis
EHR	Electronic Health Records
CNS	Central Nervous System
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
FLAIR	Fluid Attenuated Inverse Recovery
CNN	Convolutional neural Network
MLP	Multi-Layer Perceptron
PR AUC	Precision-Recall Area Under Curve
ROC AUC	Receiver Operating Characteristic Area Under Curve
GBCA	Gadolinium-based Contrasting Agent

Contribution of Authors

This thesis presents an adaptation of gradient-based meta-learning for fusion of clinical and medical imaging data in deep learning algorithms to improve medical imaging interpretation. The details of the approach are presented in Chapter 3. All of the chapters in this thesis are solely my work, with guidance by my supervisor, Dr. Tal Arbel, lab members Eric Zimmermann and Justin Szeto, and affiliate lab member Dr. Behrooz Mahasseni.

Chapter 1

Introduction

Advances in modern medicine has led to adoption of various medical imaging sensors that have tremendously improved clinical decision making. In particular, technologies such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) allow to non-invasively view complex internal structures of the body and therefore diagnose, monitor, and treat medical conditions. However, review and interpretation of high-dimensional medical imaging data (e.g. high-resolution 3D MRI scans containing millions of voxels) is a time-consuming process requiring a trained expert and is prone to manual errors. With the ever growing workload of radiological imaging exams, radiologists may need to interpret over 900 images in a typical 8 hour workday, leading to fatigue and an increased error rate in image interpretation [89]. To help ease the workload through automation, numerous machine learning methods have been adapted for a number of medical imaging tasks such as segmentation, diagnosis, and prognosis prediction. In particular, deep learning has been successfully used for pathology and structure segmentation [68, 78, 95, 163], disease diagnosis [53, 62, 157], clinical marker regression [90], and prediction of future disease flow [25, 120, 121] using high-dimensional medical imaging data (e.g. MRI, CT, PET).

In medical practice, correct interpretation and classification of organ and pathology structures (e.g. lesions) within complex medical images can often benefit from knowl-

edge of clinical context of the patient. If given only the imaging data, different structures can look similar and create uncertainty with their identification which can subsequently lead to incorrect clinical decisions. Clinical context is information that is not explicitly related to the imaging data and is gathered through non-imaging sensors (e.g. blood test results). For example, imaging features in chest X-rays resembling pneumonia (lung infection) could be attributed to a number of other conditions (e.g. lung cancer), however, presence of clinical context can confirm accurate pneumonia diagnosis if the patient also shows signs of immune response provided by context from clinical and laboratory data (e.g. presence of fever, elevated white blood cell count) [52]. Physicians often rely on clinical information from electronic health record (EHR) data to ensure effective and accurate diagnosis and treatment decisions [10]. EHR data includes information about demographics (e.g. age, sex), laboratory test results (e.g. urinalysis), patient history, disease diagnosis, and other clinical information. In one study, radiologists managed to diagnose additional causes of abdominal pain when interpreting CT scans, with the authors claiming how simple patient questionnaires can increase diagnostic yield by providing relevant clinical history to radiologists [21]. We hypothesize that using clinical information in combination with medical imaging data can lead to better performing and more accurate deep learning models by reducing uncertainties because of additional information provided by clinical data that is lacking in imaging data. One of the ways to make use of both imaging and clinical information is to look at methods proposed by multimodal deep learning techniques.

1.1 Multimodal Deep Learning

When a sensor observes some natural phenomenon (e.g. a microphone recording speech), the way the data is recorded, stored, and interpreted is referred to as a data *modality* [110]. In the context of digital medicine, the state of the patient can be observed and recorded through a variety of sensors, for example imaging technologies (e.g. MRI, CT, PET), lab-

oratory tests (e.g. blood tests, urinalysis, EHR), natural language records (e.g. patient questionnaires), and vital signal recording devices (e.g. ECG, EEG) that make up a clinical record. Digitally, these observations are stored in different ways depending on the dimensionality of the captured data, that is, how much storage is needed for a single data sample. For example, high-dimensional data such as a single high-resolution 3D MR image can contain a million voxels each containing an intensity value, meanwhile, low-dimensional tabular data such as age, blood oxygenation levels, and sex can be represented by a single value. As such, different data types (images, tabular data, natural language, speech) can be considered as different *modalities*. Furthermore, within medical imaging itself, different types of technologies (PET, MRI, CT, ultrasound) can be considered as their own modalities given that they capture information in a different manner [49].

In deep learning, multimodal fusion describes methods and models that aim to utilize complimentary information (related to a target task) from multiple data modalities in order to improve performance over single modality models [110]. In this thesis, we focus on multimodal fusion techniques that combine high-dimensional imaging data with other types of lower-dimensional clinical data (we often refer to this type of data as *tabular*). Within this context, multimodal fusion can be categorized into early fusion, joint fusion, and late fusion strategies [52, 110, 132], depending on how modalities are combined using deep learning models. Early fusion methods aim to combine data modalities at the raw data level prior to their use by a DL model. Joint fusion methods perform feature-level fusion, where features are first extracted from each modality (e.g. by another DL model) prior to their fusion. Late fusion methods perform decision-level fusion where each modality is used separately to predict a given task, at which point all predictions are aggregated to form a final consensus. Multimodal deep learning methods have found many applications in computer vision and natural language processing, including in domains such as visual question-answering [81, 124, 131, 154], and autonomous driving [18, 48]. Multimodal deep learning also saw successful applications in medical imag-

ing including lesion detection [**?**, 54, 152], disease diagnosis [93, 106, 135], and prediction of disease trajectory [111, 145, 158].

Looking at diseases such as Multiple Sclerosis (MS) and Alzheimer's disease (AD), correct analysis and interpretation of complex medical images like brain MRI are crucial for diagnosis and treatment planning. Given the difficulty of identifying various brain structures and pathologies from MRI-only information due to close visual similarities between them (e.g. 1.1), multimodal deep learning models that can utilize both imaging and clinical context information are of particular interest to aid in correct interpretation of MRI data in both MS and AD.

1.2 Multiple Sclerosis

Multiple Sclerosis is an inflammatory disease that is one of the most common causes of neurological disability in young adults [67]. MS acts as an autoimmune condition where the immune system attacks the myelin sheath, the tissue that surrounds the nerves in the brain, potentially causing damage to the nerve and causing transmitted messages to be disrupted [30]. While symptoms vary among patients, some of the main symptoms of MS include fatigue, muscle weakness, poor motor control, numbness in various parts of the body, vision problems, and difficulty with cognitive tasks such as thinking and learning [33]. Given that globally the average age of patients diagnosed with MS is 32, MS impacts adults in their productive years and severely limits their abilities and quality-of-life [30]. As of September 2020, there are 2.3 million people worldwide with MS, with two to three times as many females as males [30]. Canada's population is of special interest, as it has one of the highest rates of MS in the world at 1 in 400 [30]. Worryingly, both the global and Canadian rates of MS have been observed to increase since 2013 with no signs of stopping. Given the above, MS presents a significant threat to Canadian society due to it's debilitating symptoms as well as potential economic burden by targeting younger adults [94].

There are two main types of MS disease (often referred to as *phenotype*): relapsing-remitting, and progressive. Approximately 80-90% of MS patients are typically diagnosed with relapsing-remitting type (RRMS) [30]. This stage is categorized by sudden episodes (*relapses*) of worsening symptoms that disappear afterwards. A relapse can last from a few days to a few months. Periods between worsening symptoms are called *remission* periods. The other type of MS is progressive MS, which can be broken down into primary progressive (PPMS) and secondary progressive (SPMS) stages. During progressive MS, symptoms worsen overtime without remission periods, and this affects approximately 20% of MS patients. Meanwhile, a large portion of RRMS patients often progress into SPMS stage, which manifests as gradual worsening of MS symptoms over many years but without obvious relapse episodes [30].

The diagnosis of MS disease is done through the evidence of one or more of the following: 1) chronic inflammation of the central nervous system (CNS), 2) at least two different relapses, and 3) presence of at least two different lesions (new lesions) in the white matter of CNS [41]. Detection of brain lesion presence is done through acquisition and analysis of MRI sequences, which typically include T1-weighted, T2-weighted, Proton Density (PD), and Fluid Attenuated Inverse Recovery (FLAIR) [116]. This makes MRI one of the primary diagnostic tools for monitoring MS.

In order to investigate disease progression and pathology as well as monitor treatment effects, brain lesion activity is monitored with the use of MRI. The presence of Gadolinium-enhanced (Gad) lesions as well as new and enlarging T2 (NE-T2) lesions in MRI taken at sequential time points (e.g. images taken six months apart) act as surrogates for MS disease worsening (e.g. more relapses). In the context of RRMS patients, treatments are available that suppress new lesional activity. When a patient is on a treatment, the presence of Gad and NE-T2 lesions is used as an indication of low treatment efficacy. Furthermore, the number of NE-T2 and Gad lesions is also used as an endpoint in clinical trials during development of new treatments.

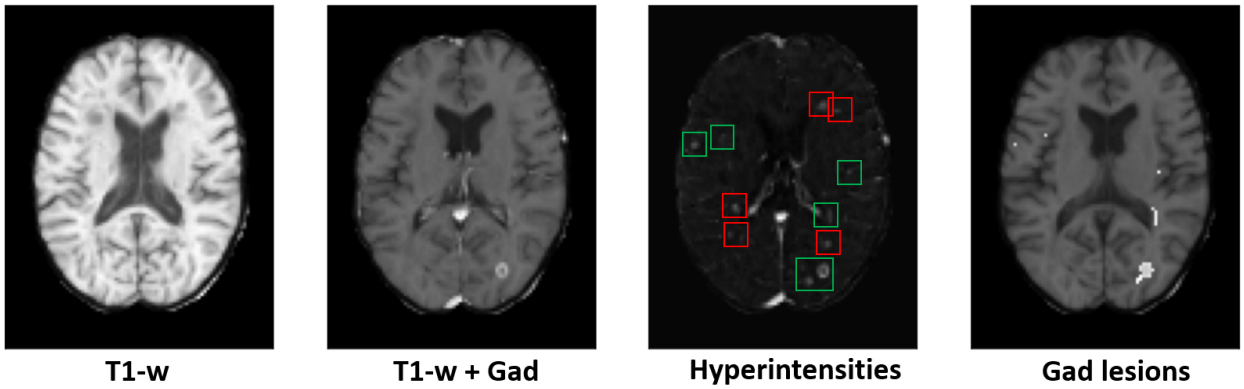

Figure 1.1: Example MRI of a patient with Gad lesions. Left to right: T1-w, T1-w with Gadolinium contrast, T1-w difference image with visible hyperintensities, T1-w with expert-annotated Gad lesions. Green boxes show hyperintensities which are labeled as Gad lesions, while red boxes show non-lesional hyperintensities. The rightmost image shows true Gad lesions overlapped on a T1-w (MRI pixel intensity values were modified to allow for better lesion visualization). Note how not all visible hyperintensities are real Gad lesions, making it hard to discern from image information alone.

To identify Gad lesions, Gadolinium-based contrast agents are administered to patients to boost visibility and delineation of new lesions in MRI scans of MS patients [33]. During an MS relapse in RRMS patients, Gadolinium contrast agent is able to pass through the blood-brain barrier and into the newly formed MS lesions [116], which become visible as high-intensity areas (hyperintensities) on post-contrast MRI scans (Figure 1.1. One major problem of identifying Gad lesions is not all hyperintensities visible in the MRI scan are Gad lesions. As seen in Figure 1.1, difference image between contrasted and non-contrasted MRI has several hyperintensity regions identified as Gad lesions (green boxes) as well several that are not (red boxes). As such, identifying which hyperintesities are real lesions solely from the MRI information is a difficult task, which can be helped through inclusion of additional clinical information.

The presence of new and enlarging (NE) T2 lesions are determined by comparing two temporally-sequential T2 MRI scans (typically at the first and the next patient visit to the clinic) and determining the number of new lesions or existing ones that have been en-

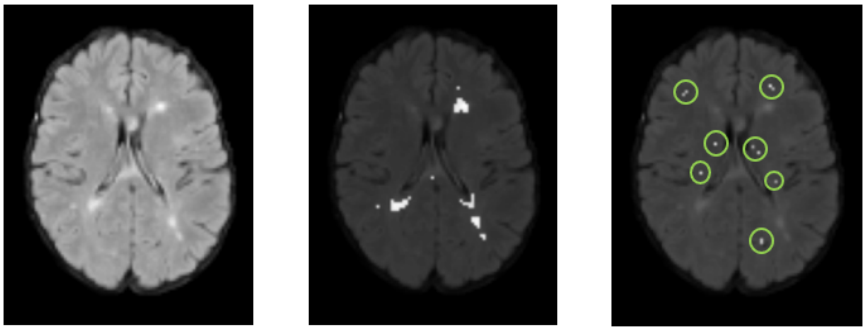

Figure 1.2: Visualization of FLAIR MRI, T2, and NE-T2 lesions. Left to right: FLAIR MRI at baseline time point, T2 lesion hypterintensities overlapped on FLAIR scan at baseline time point, NE-T2 lesion map from 48 weeks later overlapped onto baseline time point FLAIR scan. Green circles highlight locations of NE-T2 lesions. For our purposes, this patient is considered *active* as there are more than three NE-T2 lesions present [33].

larged [118] (shown in Figure 1.2). Similarly to Gad lesions, T2 and NE-T2 lesions are typically identified as hyperintensity areas on T2-w or FLAIR MRI scans [30]. Given that there are treatments that can suppress development of NE-T2 lesions, prediction of future NE-T2 lesion activity can serve as valuable surrogate marker of treatment efficacy. There has been published research in developing methods for the detection of NE-T2 lesions [22, 121] and binary classification of future disease activity [120] from MRI data. However, these methods lack the clinical context of the patient such as clinical markers or demographics information that can influence development of NE-T2 lesions (e.g. disease stage of the patient). Inclusion of additional clinical information along with MRI for prediction of future disease activity can be of benefit as shown by one study where use of both MRI and clinical data showed improved classification accuracy over MRI-only methods [25].

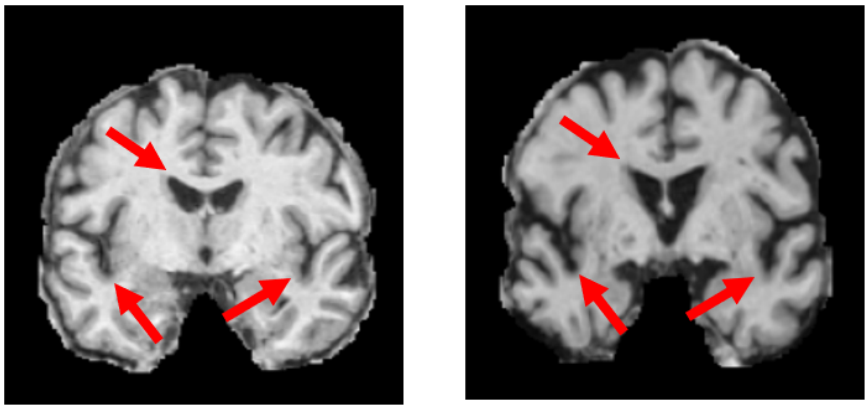

Figure 1.3: Coronal axis view of T1-w MRI scan of a cognitively normal (left) patient and one diagnosed with Alzheimer's disease (right). Both patients are female and are approximately 75 years old. Red arrows indicate differences in brain volume loss due to disease factors. Images taken from the ADNI dataset.

1.3 Alzheimer's Disease

Alzheimer's disease is a neurological disorder that causes brain atrophy and is very common in elderly patients (visualized in Figure 1.3). It is the most common cause of dementia leading to severe memory loss and decline in thinking and behavioral skills of the patient. One of the major *biomarkers* (indicators that are correlated with disease progression) for severity of AD is reduction in hippocampal volume that is visible and measured from a single time point T1-weighted MRI scan [34], however, this is not sufficient to make a diagnosis as changes in hippocampal volume can also be attributed to normal age-related changes in the brain [12]. As such, information from multiple sources (clinical tests, patient questionnaires, brain imaging) is needed for correct diagnosis. In fact, AD is diagnosed through a series of tests such as mental status and neuropsychological tests, laboratory tests, and brain imaging examinations with MRI, CT, and PET [12]. Similarly, proposed machine learning methods using a combination of imaging and clinical data have shown to perform well in classification tasks for AD diagnosis [7, 15, 135, 158],

demonstrating better performance of multimodal methods over methods using only one type of data [106].

1.4 Transfer Learning

One of the primary challenges with multimodal deep learning is determining the best method to fuse different modalities when there are large differences in dimensionality between different types of data (e.g. images vs. tabular data). For example, autonomous vehicles make use of imaging sensors (camera, LiDAR) and vehicle sensors (e.g. ultrasound, speed) that produce 3D and 1D data respectively [103]. This challenge is even more prevalent within medical imaging when attempting to use EHR data, where high-dimensional imaging data (e.g. MRI, CT) are combined with tabular data that are often just a single value (e.g. laboratory results, clinical scores). Furthermore, heterogeneous nature of how data are stored and represented between modalities often means that raw data samples require some sort of preprocessing prior to fusion in order to extract useful information [11,70]. For example, when finding ways to combine imaging and other types data of much lower dimensionality, convolutional neural networks (CNN) are commonly used in multimodal methods to preprocess raw imaging data in order to learn lower dimensional features that can then be used for fusion with non-imaging low-dimensional data [52].

While use of CNN models is a popular approach at extracting useful information from imaging data, they have also been known to learn spurious features that are predictive of the target label that are not inherently relevant to the problem (e.g. focusing on background when trained for object detection) [149]. In one example, researchers found that a CNN model trained for pneumonia detection from 2D chest X-rays learned to use metal tokens seen in training images as one of the predictive features, and as a result, performed poorly when used with an unseen real-world dataset [157]. Another study [8] found that CNN models trained for skin lesion classification were unable to capture

clinically-meaningful information and relied on obscure visual artifacts instead. Diagnosing whether the model learned spurious features is a difficult task, and even more so in medical imaging applications due to complexity of data and expert knowledge required to discern what constitutes a spurious feature [8, 37]. We hypothesize that finding ways to help a CNN model to learn imaging features that are truly relevant to clinical information can result in improvement of model performance on related clinical tasks (e.g. disease diagnosis).

One way to guide what features are learned by the imaging model is by adjusting the learned objective, that is, what is being optimized. For example, in contrasting self-supervised learning methods [57] the imaging models are pretrained to learn underlying image structures by comparing an image to a heavily augmented version of itself. Once such features are learned, the pretrained model parameters can be used for finetuning on a desired task (e.g. object classification), often showing impressive results while utilizing magnitudes-less of labeled data compared to fully supervised methods. This example is just one of many methods out there that fall under the *transfer learning* category. In short, transfer learning is an approach to use learned knowledge from one domain in order to perform better on a different domain [72]. The main assumption is that the knowledge learned during pretraining stage will be of benefit to learning the desired target task or dataset [167]. One popular transfer learning method in medical imaging applications [24, 25, 73, 136, 153, 158] is multi-task learning [117], with the idea that learning several *related* tasks simultaneously by a single model can capture intertask differences and be beneficial to model performance. However, this approach creates additional complexity in the design of the optimization objective (e.g. which tasks are more important) as well as the learning procedure in general, given that multiple optimizations are happening at once. In contrast, another transfer learning method called *meta-learning* [50] aims to achieve a similar goal as multi-task learning but instead aims to use full representational power of the model to focus on learning each task individually (explained further in the next section). We hypothesize that a meta-learning approach to learning multiple relevant

tasks at once can help guide a CNN model to learn more meaningful and relevant imaging representations for the clinical context.

1.5 Meta-learning

Meta-learning is a subset of transfer learning focusing on *learning to learn*, that is, observing how algorithms perform on a distribution of learning tasks and then learning from this experience [79]. Meta-learning methods aim to train models that capture prior knowledge (also referred to as *inductive bias*) from a distribution of related tasks. This inductive bias (contained in the trained model parameters) can then be used to quickly adapt the model and still perform well on a previously unseen task [50]. The intuition behind meta-learning is to treat entire tasks as training examples in order to generalize well over a distribution of tasks, compared to conventional ML where model is trained using multiple data instances to generalize across the dataset [31]. Meta-learning is also a common technique for few-show learning, tackling the problem of training models with little available data [141] which is prevalent in medical imaging domain due to difficulty of acquisition and data labeling [75]. Meta-learning approach allows pretrained models to quickly learn a new task using only a few data samples or training cycles, and has seen many applications in few-shot detection, segmentation, and image generation [42, 102, 139, 151]. In medical imaging applications, meta-learning has been successfully used for segmentation tasks where the model that was pretrained to segment a set of organs is able to generalize well and perform accurate segmentation on an unseen organ from a few training samples [29]. Similarly, another study used meta-learning approach to train DL models in a low-data regime by first training a CNN to learn a set of common disease classification tasks and then adapt this model to successfully diagnose rare skin diseases from only a several training samples [79].

In this thesis, we conjecture that deep learning model outcomes based on medical images alone will show improvement should addition clinical information (e.g. demograph-

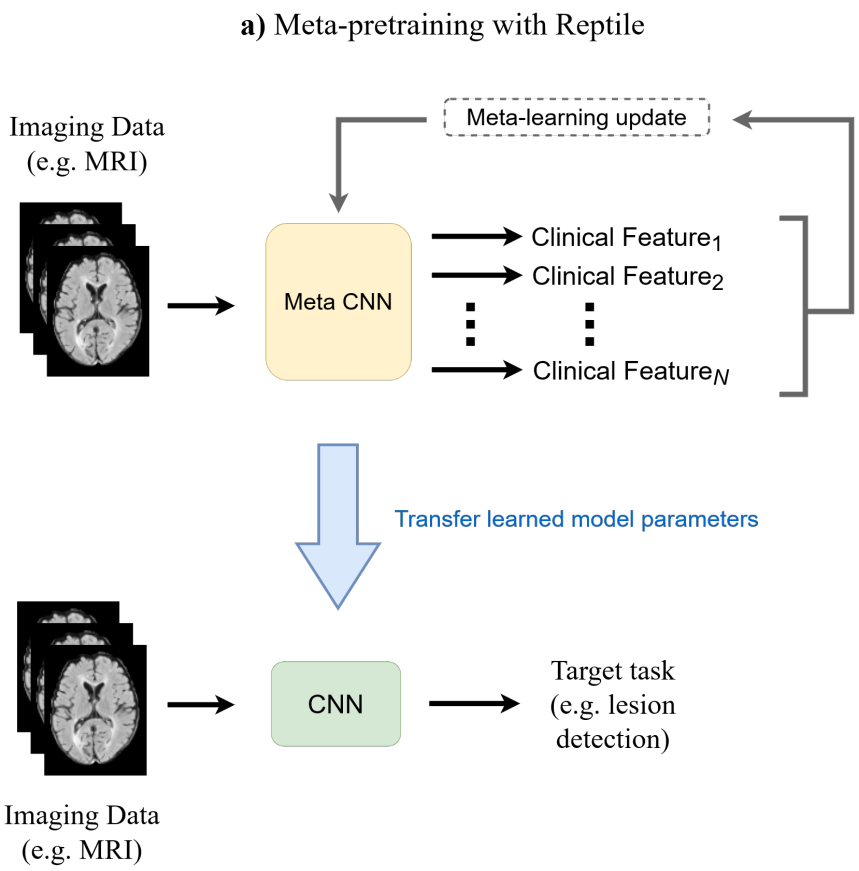

Figure 1.4: Simplified overview of the proposed meta-learning method for embedding clinical context into learned imaging features. Top (**a**): a CNN model is pretrained using a Reptile meta-learning algorithm to perform well at predicting a set of clinical features from MRI data. Bottom (**b**): the pretrained weights are used as parameter initialization to finetune the CNN on the desired target task using the same MRI data.

ics, disease state) be provided to the network via meta learning, with the assumption that each data modality (e.g. MRI, laboratory test results, doctor reports) carries unique information about the clinical context. This is accomplished by adapting a gradient-based meta-learning algorithm called Reptile [97] for pretraining a CNN model, and tested using MS and AD medical datasets. The intuition is to guide a CNN model to learn imaging features from MRI data that can capture relevant clinical context and carry complimentary information to help improve performance on a desired related clinical task (overview

in Figure 1.4). First, a CNN model is pretrained on MRI data to independently classify or regress a number of relevant clinical features (e.g. cognitive test scores, disease stage, lesion volume) using a meta-learning scheme. The pretrained CNN model is then fine-tuned on a desired target task using the same MRI data as input. For evaluation purposes, we apply the above method to train deep learning models on three example target tasks: 1) detection of Gad lesions in MS, 2) prediction of future NE-T2 lesion activity in MS, and 3) estimation of ADAS-13 and MMSE cognitive scores in AD patients using both MRI and tabular clinical data. Through experimentation, we compare our method to single modality (*unimodal*) methods using only imaging or only clinical data as well as existing multimodal and transfer larning methods in medical imaging. Additionally, we perform a study on the selection and impact of individual clinical markers on the target task performance of our proposed approach. Lastly, given that meta-learning methods have been previously used for training models in low-data regimes [79, 162] and the prevalence of the data availability problem in medical imaging [143], we also investigate the effect of dataset size on the performance of our proposed approach.

1.6 Contributions of Thesis

The work presented in this thesis demonstrates how meta-learning approach can be used to enhance the quality of imaging features in deep learning algorithms through provision of clinical information, and consequently, improve performance of the imaging DL models on desired target tasks. Through several experiments, we demonstrate the following:

1. **Present an adaptation of meta-learning approach for enhancing performance of image-based deep learning algorithms on medical interpretation tasks.** We hypothesize that correct interpretation of medical images with deep learning models can benefit from addition of tabular clinical context. This thesis presents a meta-learning approach to infuse knowledge of clinical information directly into the imaging features extracted by a convolutional neural network. These pretrained features

are then used to finetune the network on a related target task, and show improved classification performance over methods using imaging-only data as well as competitive performance against existing multimodal and transfer learning methods in the medical imaging domain. We experiment on a real-world MS and AD datasets in our experiments to show examples of real-world applications on tasks of lesion presence detection, prediction of future lesion activity, and regression of clinical scores.

2. **Quantitative analysis of the impact of clinical feature selection and dataset size on performance of the proposed method on the target tasks.** Through experimentation, we explore the impact of individual clinical features on the meta-learning process and model performance, and provide a guidelines on their selection given the target task. Furthermore, given the common problem with labelled data availability in medical imaging, existing meta-learning methods have been used to tackle this problem in this domain. Similarly, we conduct a short exploratory study to investigate how well our proposed approach performs across various data regimes.

1.7 Thesis Overview

This thesis presents a meta-learning approach to fusion of clinical and medical imaging information in deep learning algorithms for detection and prediction of lesion activity in MS patients and regression of clinical scores in AD patients. The thesis is structured as follows.

Chapter 2 provides background knowledge on design, training procedures, and evaluation metrics of deep learning algorithms. We then provide definition of data modality, multimodal fusion in deep learning, and present a literature review of existing methods and their applications in computer vision as well as the medical imaging domain. The concepts of transfer learning, multi-task learning, and meta-learning are then discussed

along with background on their existing applications in medical imaging and computer vision.

Chapter 3 presents our proposed approach of utilizing Reptile meta-learning method for combining imaging and clinical data to improve performance of image-based DL methods. We present the original Reptile method in detail and describe how it is adapted for fusion of clinical and medical imaging information. We then discuss the thought process and considerations behind selection of clinical features as well as Reptile-specific hyperparameters and their impact on model performance.

Chapter 4 starts by covering implementation details of our proposed method and experimentation pipeline. In this chapter, we introduce example problems used for evaluating the proposed method, which are 1) detection of Gad lesion presence, 2) prediction of future NE-T2 lesion activity, and 3) regression of ADAS-13 and MMSE cognitive scores, all from baseline brain MRI scans and supporting clinical data. Compared to using imaging-only DL methods, the above example tasks were chosen as we believe they can show benefits of using both imaging and clinical data to provide clinical context to the image-based deep learning model and help discern ambiguities that are present when using imaging-only data (described further in Sections 1.2, 1.3, and in Chapter 4.3). In the case of task #1, using only MRI information is difficult because Gad lesions can be mistaken for unrelated hyperintensities (Figure 1.1). For tasks #2 and #3, providing DL model with only MRI information presents only one data point for prediction of future NE-T2 lesion activity and estimation of the cognitive scores respectively, and is insufficient to fully capture the clinical context of the patient and make accurate predictions. The chapter then presents the details about the MS and AD datasets along with the relevant statistics, evaluation metrics, and the data preprocessing steps we used. Additionally, we list and describe selection of clinical features and MRI sequences along with rationale behind the choices. Lastly, we describe in detail unimodal, multimodal, and transfer learning baseline methods used for comparison with our proposed approach later in Chapter 5.

Chapter 5 presents experimental results of the proposed meta-learning approach and baseline methods on the three selected target problems. The proposed approach showed general improvement of 1-3% across metrics over imaging-only method for the Gad lesion presence detection (task #1) and prediction of future NE-T2 lesion activity (task #2), as well as achieved lower root mean squared error on estimation of ADAS-13 and MMSE cognitive scores (task #3). Additionally, meta-learning approach also achieved marginally better metric performance than all other baseline methods for tasks #1 and #3, but not with task #2 where a modality bias was identified. Ablation experiments on selection of supporting clinical data confirmed our assumption that clinical features which are closely related to both the target task and the visible structures in the MRI generally achieved the best performance on the target task, demonstrated by experimental results on tasks #1 and #3. while it is unknown whether the above findings are statistically significant (due to computational and time resources required, focus instead on thorough experimentation and use of cross validation), we believe the extensive experimental results are sufficient to give insight on the trends discussed above. Lastly, a short investigation into the effects of dataset size on model performance showed that meta-learning approach can perform competitively compared to other baseline methods across varying data regimes. Specifically, experimental results on task #1 showed meta-learning approach achieving on-par or better ROC AUC and PR AUC metrics compared to the next best performing method across all but the lowest data regime. We conclude the chapter with a discussion of the limitations of our proposed approach.

Chapter 6 concludes the thesis by summarizing the findings and key insights presented in Chapter 5, as well as provides some thoughts on future work.

Chapter 2

Background and Related Works

This chapter provides a thorough overview of deep learning fundamentals, multimodal fusion techniques, concepts of transfer learning and multi-task learning, as well as meta-learning methods in medical imaging. First, relevant background knowledge about deep learning concepts, training procedures, and evaluation metrics is presented. It is followed by a review of multimodal fusion methods, examples of use, and their application within the medical domain. Lastly, the concepts of transfer learning, multi-task learning, and meta-learning, are presented along with their various applications in computer vision and medical domain. Information presented in this chapter provides the required knowledge foundation for the following chapters.

2.1 Deep Learning

Deep learning (DL) is a subset of machine learning (ML) that focuses on training and using artificial neural network (ANN) algorithms to automatically learn useful features from input data and perform desired tasks [72]. In traditional ML, data preprocessing involves manual creation of useful features that can be then passed to the ML algorithm, which lead to many years of research into feature engineering [46]. With deep learning, this process of feature extraction from data is automated during training of DL models,

where the model automatically selects what features to extract in order to learn the task at hand. In recent years, DL methods have exploded in popularity by reaching state-of-the-art performance with applications in computer vision (CV), natural language processing (NLP), graph learning, and many other domains.

2.1.1 Convolutional Neural Networks

In computer vision applications of DL, Convolutional Neural Networks (CNN) are one of the most popular types of NNs, and have been designed to efficiently deal with structured data (matrix-like, e.g. images) as they are able to learn spatial hierarchical patterns in structured data [3]. CNNs are made up of convolution layers, non-linearity activation functions, pooling layers, and fully-connected layers (e.g. multilayer perceptorns, or MLP). Convolution layers (along with activation functions) are used for feature extraction, pooling layers are used for downsampling, and fully connected layers are used to consume extracted features to perform the desired task (e.g. classification).

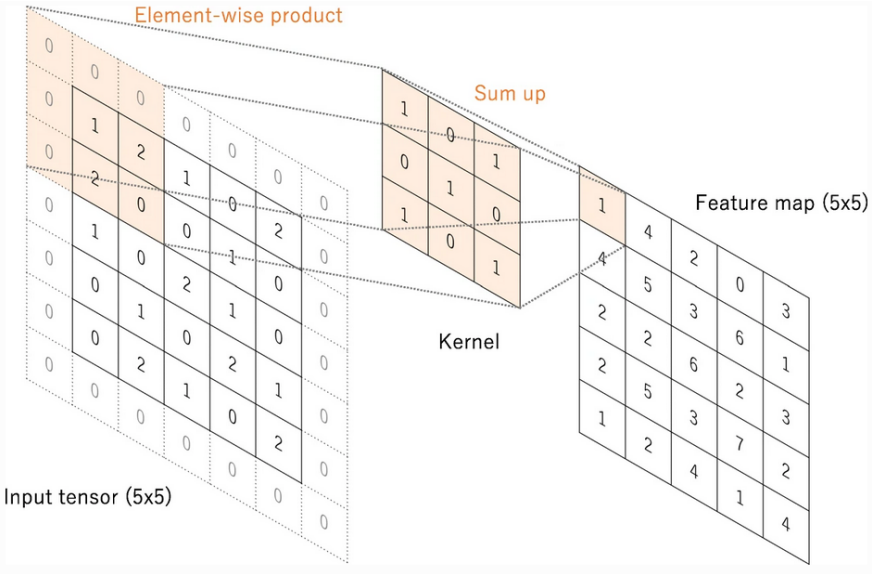

Figure 2.1: Visualization of convolution operation in CNN on a 5x5 input matrix using a 3x3 kernel with zero padding and stride of 1. Courtesy of [146], used with permission.

The core component of CNNs are convolutional layers. Typically paired up with activation and normalization operations, convolutional layers perform convolution operation on the input data matrix (pictured in Figure 2.1). During a convolution operation, a matrix of learned parameters (*kernel*) performs element-wise product and sum in a sliding-window manner across the input data matrix to produce a feature map. The step of the kernel during sliding-window operation is referred to as *stride*. In order to allow CNN to model non-linear relations, feature maps are passed through a non-linear activation functions such as ReLU, sigmoid, tanh, and others [20, 72]. Afterwards, a pooling layer such as Maxpool or Averagepool is used to downsample feature maps by selecting either the maximum or the average value within the patch area for the output (visualized in Figure 2.2). Pooling operations are used to reduce number of learnable parameters and introduce translation invariance.

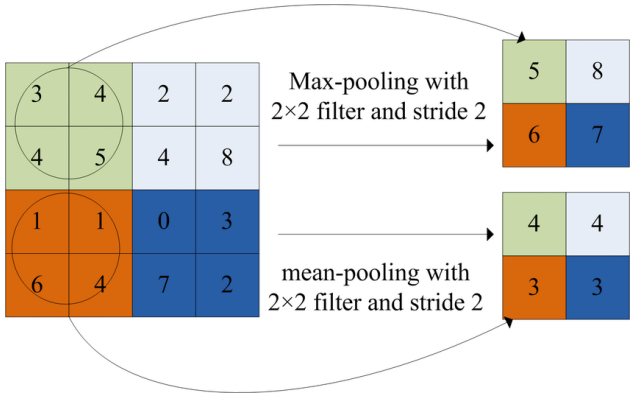

Figure 2.2: Visualization of maxpool and meanpool (identical to averagepool) operations with 2x2 filter size and stride of 2. Courtesy of [86], used with permission.

2.1.2 Training Neural Networks

When training NN models, a given dataset is typically split into training, validation, and testing sets. Data samples contained in the training set are used for training the model, meanwhile the validation set is used to measure how the model performs on unseen data

during the training process. The test set is used at the end of the training to evaluate generalization performance of the trained model on unseen data.

In a supervised learning setup, neural network models are trained by updating the model parameters (also known as *weights*) in an iterative manner to reduce the difference between predicted output and ground-truth labels. Using the training set, this is done by feeding data samples into the model and comparing the predicted output to the corresponding ground-truth label using a selected metric, called a *loss function*. Loss functions vary widely depending on the learning objective, however, all have one necessary condition: a loss function must be smooth differentiable in order to ensure a useful gradient of the loss can be computed. Some common ones include binary cross entropy loss for binary classification tasks, mean squared error for regression tasks, and cross entropy loss for multi-class classification.

An optimization function is used to determine how should the model parameters change in order to minimize the loss function, and as such, bringing the predicted model outcome closer to the ground truth. There are a number of different options for optimization functions, including Stochastic Gradient Descent, AdaM, RMSProp, and Adagrad [66,155]. Gradient descent uses the derivative of the loss function with regards to the model parameters to update said parameters in the direction negative to the gradient, and thus minimizing the loss [4]. The magnitude by which the model weights are changed is called the *learning rate*, and it controls how "fast" the learning takes place (e.g. higher learning rate means greater change in weights). The gradient is then propagated through all of the model parameters using the back-propagation algorithm, and each parameter is updated using the gradient descent scheme [47] (visualized in Figure 2.4).

The act of passing a data sample through the model and getting a prediction is referred to as a *forward pass*, while the act of updating model parameters given the calculated loss between prediction and ground-truth is known as *backward pass*. A training *iteration* is a single instance of forward-backward pass, where the model parameters have been updated a single time. During the training process, there are many *iterations* and thus

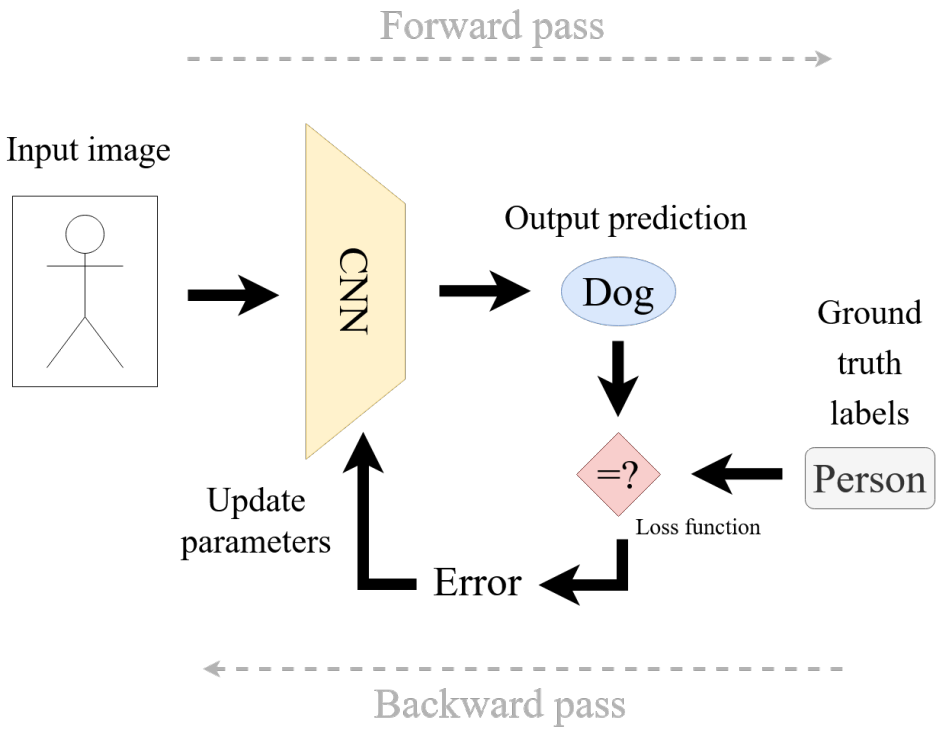

Figure 2.3: Training of a CNN model using supervised learning learning approach. During forward pass, input data are passed through the model to obtain a prediction. The prediction is compared to the ground truth label using the loss function to obtain an error metric, which is then used to update model parameters using an optimization algorithm and backpropagation.

many model parameter updates, as the model consumes entirety of the training set. One cycle of the model consuming all of the training set samples is called an *epoch*, and is made up of many iterations. Models are trained over many epochs, with the number of epochs being one of the hyperparameters selected by the user.

There are two main aspects to consider during the model training stage: training and validation losses. Monitoring training loss provides feedback on how well the model is learning, that is, model predictions get closer to the ground truth labels and the loss should decrease as the model completes more training epochs. In the meantime, we also want to make sure that the model is not *overfitting* to the training data and still provides good performance on unseen samples. After every training epoch, all samples in the validation set are passed through the model and the validation loss is calculated. Once the

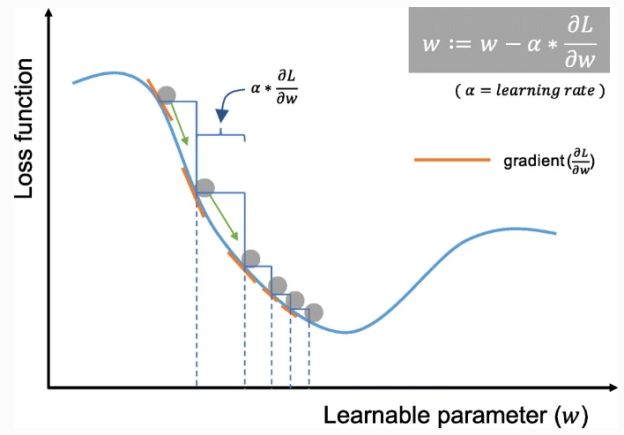

Figure 2.4: Visualization of gradient descent optimization procedure on the learning parameter w. Courtesy of [146], used with permission.

validation loss plateaus or begins increasing (Figure 2.5), training should be stopped even if training loss continues to decrease as this is a sign of overfitting (Figure 2.5). To combat overfitting and improve generalization, various regularization techniques are used, for example data augmentation and dropout. Data augmentation introduces random changes to the input data during training to introduce extra variation within the data, and by extension, artificially increase the size of the dataset. Ideally, these augmentations should be realistic and non-destructive to the features of the desired task. In computer vision domain, some of these include image augmentation operations such as rotation, flipping, minor Gaussian noise, or contrast changes [126]. Dropout is another approach to combat overfitting [130]. With dropout, random neurons in the neural network are set to zero, reducing the learning capacity of the network and approximates training large number of models with different architectures. This method has been shown to be very effective at reducing overfitting [130].

Another aspect of DL model training is hyperparameter tuning. Using the validation set performance, hyperparameters can be adjusted in order to improve performance on unseen data [148]. Hyperparameters are defined as the parameters that control the learning process and are specified by the user. These include batch size, number of epochs, amount of augmentations, learning rate, loss and optimization functions, and regulariza-

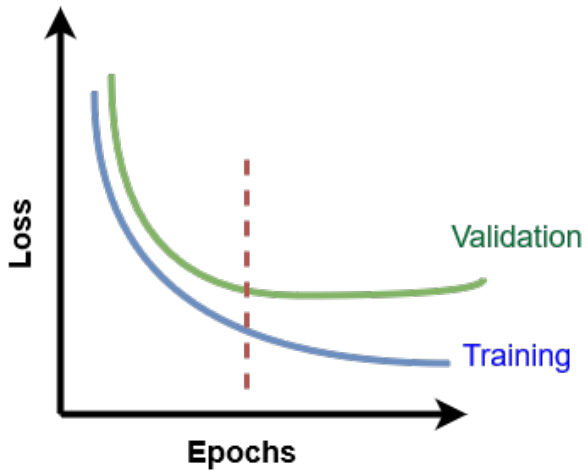

Figure 2.5: Point of overfitting (dotted line) during model training

tion strategy. Additionally, specifying model architecture parameters can be included as well, such as number layers in the network, types of normalization layers, CNN kernel size and stride, and activation functions.

2.1.3 Evaluation Metrics

While measuring training and validation losses is important to ensure proper model training, there are other metrics that are directly related to the task which are used to decide whether the model performs up to our expectations after training. For evaluating binary classification tasks, a popular approach is to a binary confusion matrix seen in Figure 2.6. This matrix contains counts of four outcomes of a binary classifier: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Combinations of these counts make up a number of useful metrics for evaluating performance of a binary classifier.

Recall, also known as Sensitivity, measures the number of correct positives out of all true positive predictions (true positives and false negatives).

$$Recall = \frac{TP}{TP + FN} \tag{2.1}$$

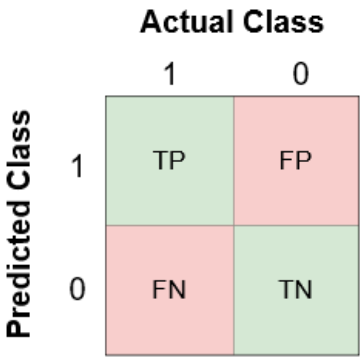

Figure 2.6: Binary confusion matrix

Precision, also known as Positive Predictive Value (PPV), measures the number of correct positives out of all true and false positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

F1-score is a harmonic mean of precision and recall, and is a useful measure when a class imbalance is present (e.g. many more samples of class 1 compared to class 2).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.3}$$

It is important to note that all of the above metrics rely on the selection of a *threshold* that will binarize model output and determine what is considered positive or negative class (e.g. with 0.5 threshold, anything 0.5 and above is considered positive class, and everything below as negative class). Given how critical threshold selection is, we make use of Precision Recall Area Under Curve (PR AUC) as well as Receiver Operating Characteristic (ROC) curve AUC. Precision-Recall curve is a plot of precision and recall against each other and ROC curve is a plot of recall vs. false positive rate (FPR), both at a wide range of binarization thresholds. ROC AUC and PR AUC are the areas under their respective curves, and are bounded between 0 and 1 range (1 being the ideal classifier). These metrics are important as they allow to compare model performance without the dependency on correct selection of the decision (binarization) threshold.

Performance evaluation of regression tasks is commonly done by measuring the error between the predicted value, $y^{predicted}$, and true value, y^{true}. One example is mean squared error (MSE) which measures the average squared error across all N data samples. A modified version of MSE is root MSE (RMSE) which takes the square root of MSE value, with the benefit of measuring the error in the same units as the predicted value.

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n^{true} - y_n^{predicted})^2 \tag{2.4}$$

2.2 Multimodal Fusion In Deep Learning

When a sensor (e.g. digital camera) observes some phenomenon in the real world, the way the sensor records and stores data that represents that event is referred to as a data *modality* [110]. Different sensors and modalities allow to capture complimentary yet non-overlapping information when recording the same event. For example, capturing a video of a barking dog records both image and sound which produces different data modalities that both independently identify the subject (a dog). In deep learning, use of multiple modalities is motivated by the assumption that there is complimentary information provided by each modality that can be used by a DL model [110]. The goal is to use this additional modality-specific information in order to improve performance on selected tasks (e.g. object classification) compared to methods that use only a single modality.

Given the differences in representation between data modalities (e.g. grid-like 2D image arrays vs. 1D time-series sound recording), one of the primary challenges multimodal fusion is how to best combine and make use of various types of data [51]. In multimodal deep learning, there are three main design choices that must be decided on prior to implementation: 1) how to fuse different modalities, 2) which modalities to use, and 3) how to deal with missing data [110]. The first design choice is how to fuse different modalities (an overview of various DL multimodal fusion methods is visualized in Figure 2.7). Also known as *data fusion*, combining raw data of several modalities early-on into a unified

feature vector prior to feeding into a DL model is known as *early fusion*. Given that fusing raw data samples can be often be challenging due to differences in dimensionality and representation (e.g. 3D image data and 1D time-series data), extraction of high-level representations (features) prior to fusion can help alleviate these issues [110]. With *joint fusion* methods, feature representations are first learned by the separate layers (e.g. CNN layers for imaging data) from the raw data. Then, these features are fused together (e.g. concatenation, pooling) to create a shared representation used by a DL model for the desired task (e.g. classification). The main difference between early and join fusion is that with joint fusion, the training loss is also used for training the feature extraction model and thus, modifying feature representation to better suit the end task. Lastly, *late fusion* (also known as *decision-level fusion*), refers to training separate models on raw data of different modalities to get predictions that are then aggregated (e.g. averaging, majority voting) to make a final decision.

The second design choice, and potentially the most important, deals with selecting which modalities to fuse. The core assumption of multimodal fusion is that different modalities can provide complimentary and useful information to the task being solved [51]. It is also possible that inclusion of some modalities can be actually detrimental to model performance, and thus, diligent feature selection is vital to ensure that the benefits of multimodal fusion are realized. Feature selection can be done manually with the prior knowledge of the user, however, there has also been research into learning the optimal selection of modalities and fusion architectures. Previous studies proposed methods such as pruning algorithms [112], genetic algorithms [125], as well as forms of reinforcement learning [168] and Bayesian optimization [109] with various degrees of success [110].

In use cases where data from multiple modalities are present, multimodal fusion has often shown to provide significant performance improvements over single modality DL methods. In the domain of human activity recognition, multimodal fusion methods combine audio, video, depth, and skeletal motion data modalities to solve problems such as action recognition [**?**, 108] and human pose estimation [114]. In another example, vi-

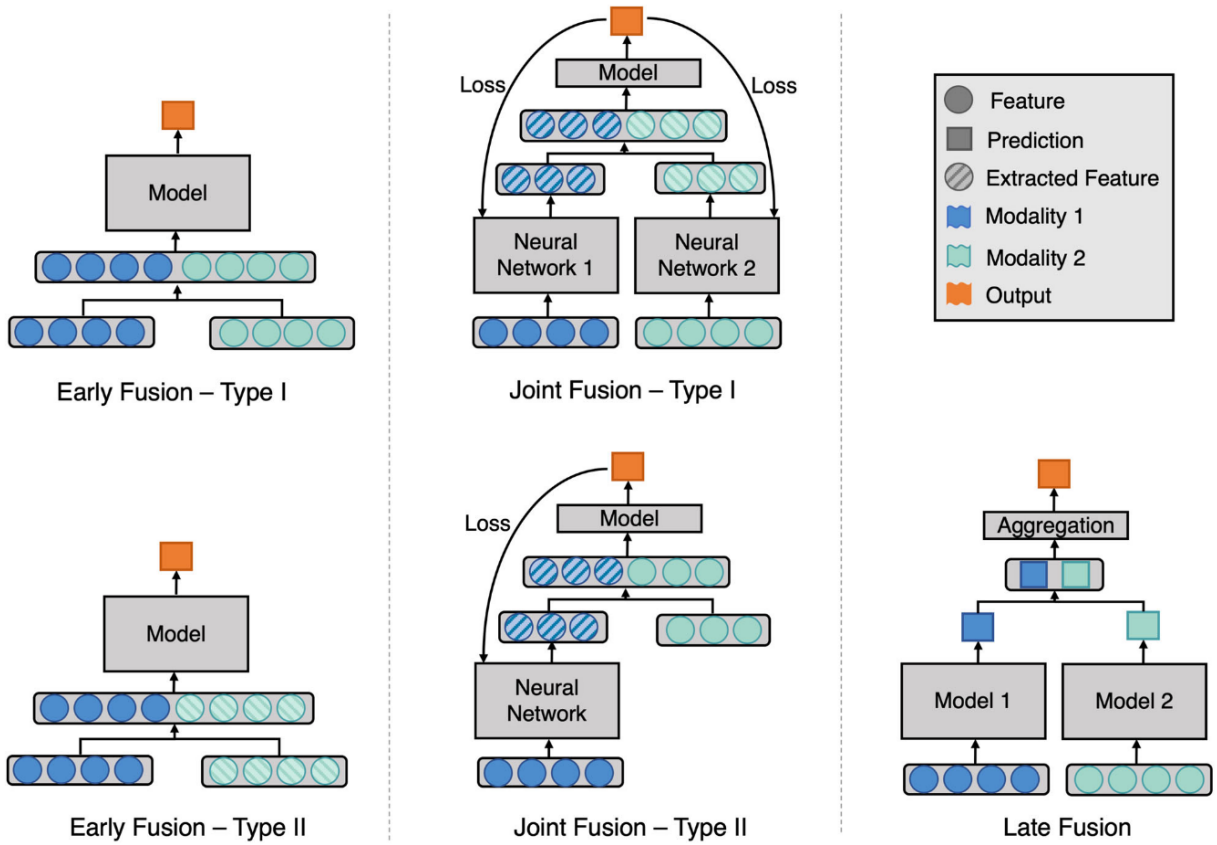

Figure 2.7: Categories of multimodal fusion methods in deep learning. Early fusion methods combines raw data (type 1) or extracted features (type 2) prior to model input. Joint fusion methods extract learned features prior to fusion and use the training loss to guide feature extraction process. Late fusion aggregates predictions from separate models using separate modalities to provide a single output. Blue and cyan circles represent raw data of separate modalities, slashes represent extracted features, and squares represent predictions. Courtesy of [51], used with permission.

sual question answering (VQA) tasks by their nature deal with multimodal data (natural language text and images), where the objective is to produce a natural language answer given an image and a related natural language question [134]. Due to dimensionality differences between modalities, a common approach in VQA problems is joint fusion [134] where learned features from text and imaging data are first extracted with the use of CNN, MLP, and autoencoder models prior to fusion and use in a downstream DL model [13, 55, 101], often reaching state-of-the-art performance [107]. In autonomous

driving applications, multimodal data are often captured from multiple sensors (LiDAR, radar, camera, proximity), giving popularity to applications of DL multimodal methods [58,96,99,144]. In one case, combining camera and LiDAR data has lead to higher accuracy in object detection tasks over LiDAR-only methods [103]. Applying a deep learning method to determine if a person is lying or telling the truth, Gogate *et al.* showed that combining audio, visual, and text data improved performance when compared to existing state-of-the-art methods, with early fusion outperforming late fusion approach [40]. There has also been extensive use of multimodal fusion for scene understanding and semantic segmentation space in computer vision domain, often reaching state-of-the-art performance [163].

2.2.1 Medical Imaging

Digitization of modern medicine generates many types of data and modalities, such as imaging data, natural language data (e.g. doctor reports), and EHR data such as clinical information (e.g. diagnosis), demographics (e.g. age, sex), and laboratory results [51]. While different types of medical images (e.g. MRI, CT, PET, X-ray) can also be considered as separate modalities due to differences in how data are represented and interpreted, we specifically focus on fusion of medical images with low-dimensional clinical data (e.g. clinical scores). From cancer risk prediction to brain lesion detection and disease diagnosis, utilizing multiple data modalities has shown to reduce visual ambiguities and improve performance of DL methods for interpretation of medical images [51]. Within the medical imaging domain, the most common multimodal fusion approach is early fusion. Early fusion methods directly concatenate clinical data with imaging features that have been extracted by using a CNN or other methods (e.g. analysis software). Many studies using early fusion methods have demonstrated improved performance in tasks like predicting symptom progression, lung cancer subtypes, and bone density estimation [7,25,54,74,93]. Applications of joint fusion typically first extract feature representations of each of the modalities (with the help of MLP or CNN layers) prior to their fusion.

These methods are not as common as early fusion, however, still show improvement over single modality methods [63, 129, 145]. Lastly, studies using late fusion techniques that aggregate predictions from unimodal models also showed improvement over single modality methods when used for diagnosis of Alzheimer's disease [106], prostate cancer [111], and pulmonary embolism [52]. While multimodal methods are a popular way of utilizing imaging and clinical data, *transfer learning* is another commonly used set of methods for learning useful information from both modalities.

2.3 Transfer Learning

One of the primary challenges to wider adoption of machine learning applications is the lack of availability of sufficient labelled training data, which is often the case in a lot of real-world problems [167]. One promising solution to tackle the lack of labelled data in one domain (e.g. task or dataset) is to exploit knowledge learned from another. The set of machine learning methods that do this are known as transfer learning (TL) [72]. The intuition is that the knowledge learned from one domain (e.g. classifying animal species) can be leveraged to improve learning performance on a *related* domain (e.g. classifying dog breeds). In practice, a model is first pretrained on a labelled dataset to learn useful representations (knowledge) from the data (captured by the trainable parameters of the model). This pretrained model is then finetuned (either entire model or select layers) on the target dataset or task, with the goal of decreasing training time and/or improving model performance on the target task [167]. One of the key considerations, aside from availability of labels, is selecting a pretraining (source) domain that is *related* to the desired (target) domain in order for knowledge transfer to be successful [167]. Otherwise, a negative transfer phenomenon can occur where transferred knowledge is "unrelated" and negatively affects performance on the desired domain [161]. In computer vision and NLP, TL methods have gained popularity due to existence of many large labelled large datasets that can be used for pretraining [167], and have been shown to have improved

generalization and training time over training networks from scratch with randomly initialized weights [26]. Aimed at tackling different problems, there are a number of variations of transfer learning techniques including domain adaptation [62], self-supervised learning [59], multi-task learning [117], and meta-learning [50]. Domain adaptation addresses problems where there is a drift in data or label distribution between source and target domains. Self-supervised learning aims to learn useful features from unlabeled or very small amounts of labeled data. Multi-task learning aims to jointly learn a set of related tasks from the same data by taking advantage of similarities and differences between the tasks. Lastly, meta-learning aims to learn over a distribution of related domains in order to generalize well to an unseen domain. In the next section, we provide further background on multi-task learning and meta-learning.

2.3.1 Multi-task Learning

Multi-task learning is a technique for training machine learning models where several related tasks are jointly learned (visualized in Figure 2.8). Compared to traditional transfer learning with sequential training (train on source then on target domain), the intuition behind multi-task learning is that by paying attention to all related tasks simultaneously the model can take advantage of intertask relevance and differences and thus, lead to better generalization and performance [117]. As seen in Figure 2.8, typically a shared representation is first learned by a backbone model (e.g. CNN layers for image data), followed by separate task-specific layers. By learning a shared representation between related tasks, the model learns an inductive bias with the assumption that the information learned for one task can also be beneficial for learning other tasks [137]. Various applications of multi-task learning have been successful in both computer vision and medical imaging domains [1, 80, 147, 153, 164]. In medical imaging, Zhang *et al.* utilized multitask learning in a multimodal setup for feature selection and successful estimation of several clinical markers for Alzheimer's disease progression [158]. In the meantime, authors of [53] used multi-task learning for joint diagnosis of several mental disorders from

functional MRI in a federated learning setup. In another study, multi-task approach to lesion segmentation from MRI resulted in improved Dice score and segmentation accuracy of small brain lesions [95]. Lastly, a multi-task approach has also been used for prediction of treatment efficacy in MS patients from brain MRI and select clinical data [25].

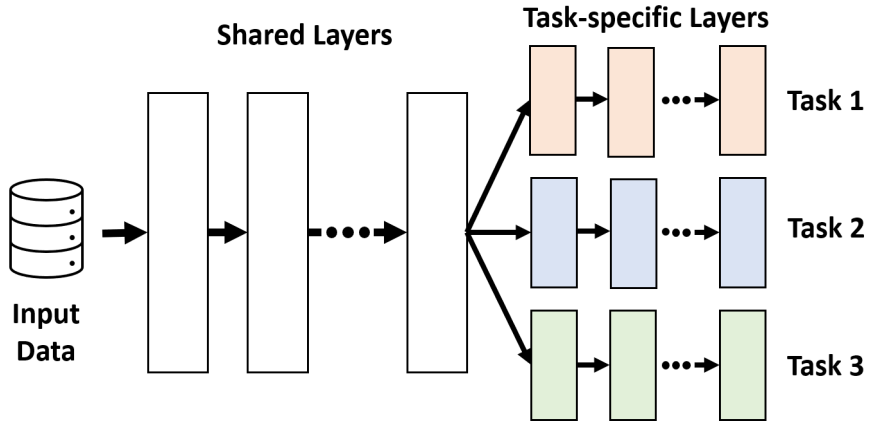

Figure 2.8: Visualization of multi-task learning scheme with hard parameter sharing which aims to jointly learn several related tasks. Typically, a shared representation of the input data are learned through a number of shared layers, which is then used by separate task-specific layers to learn each task. Training loss from each of the learned tasks is used to adjust the learning of the shared representation accordingly.

2.3.2 Meta-learning

Meta-learning is another type of transfer learning aimed at generalizing across a distribution of tasks in order to improve future learning performance on an unseen task, and has found many uses in training models with small amounts of data [50]. Compared to traditional learning schemes that aim to train models to directly solve tasks at hand, meta-learning instead aims to *learn-to-learn*, that is, to improve on the learning process itself. This is achieved through pretraining models over multiple learning episodes (rather than multiple data instances as in traditional ML) using a distribution of related tasks in order to capture inductive bias (prior information) relevant to the task distribution [50]. This learned experience can then improve future training performance and provide computa-

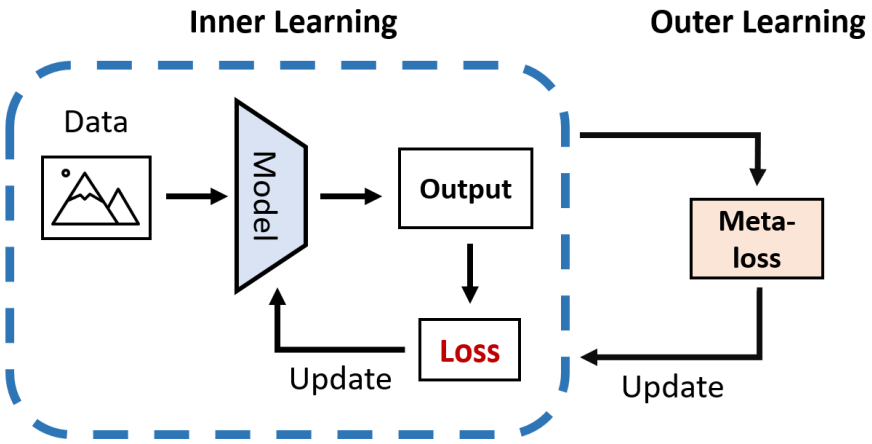

Figure 2.9: Overview of meta-learning scheme. An inner learning stage consists of training a model on a specific task. An outer learning stage adjusts the inner learning algorithm based on a specific objective.

tional benefits when the model is finetuned on an unseen related task. There are typically two phases in a meta-learning scheme: base (inner) learning and meta-learning (outer) stages [50]. During the base learning, a learning algorithm (e.g. CNN) is trained to solve a specific task (e.g. classification) defined by an objective and a dataset. During the meta-learning stage, an outer algorithm updates the inner learning algorithm to improve a certain outer objective (e.g. generalization across tasks, speed of training, etc.). For example, by using this definition, the process of hyperparameter optimization by cross-validation could be considered as an example of meta-learning. While there are various approaches to meta-learning, in this thesis we specifically focus on gradient learning methods [31,97] due to the relative simplicity of their implementation.

Gradient-based meta-learning methods work by adjusting the optimization process to achieve the meta-learning goals. One popular approach is model-agnostic meta-learning (MAML) [31], a method that aims to learn feature representations such that they are broadly suitable for a number of related tasks while remaining model-agnostic (no requirements for model architecture). The resultant model allows for fast learning on a desired domain through finetuning with a small amount of gradient updates. A short overview of how MAML works is the following. During meta-pretraining, each task

$t_i \in T = \{t_1, t_2, ...\}$ has a support data set $\mathcal{D}_i^{tr}$ and a query test set $\mathcal{D}_i^{ts}$. In a single iteration for a single task t_i, the parameters of model Θ are updated such that if Θ is finetuned on support set $\mathcal{D}_i^{tr}$ (resulting into model parameters Φ), it also performs well on the query test set $\mathcal{D}_i^{ts}$ and thus, generalizes well on both sets. The above process is repeated for all tasks, and the final gradient update to the model Θ is a combination (e.g. sum, average) of gradients across all tasks $\mathcal{T}$, resulting in model $\Theta^{'}$. The above process is visualized in Figure 2.10. One major downside of MAML is the need to compute second derivative terms (a Hessian) for every task which is computationally expensive, and gave a rise to further modifications that aimed to solve this problem [27, 97].

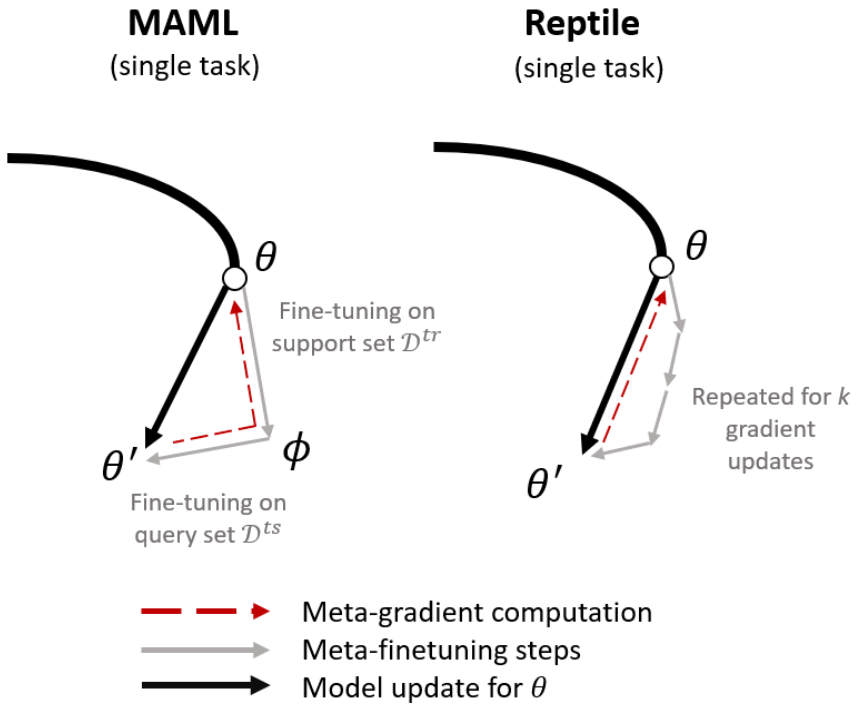

Figure 2.10: Comparison between MAML [31], and Reptile [97] meta-pretraining for a single task. MAML computes the meta-gradient through fine-tuning on the support and query sets sequentially. With Reptile, the meta-gradient is computed by multiple gradient updates on the selected task, without the need for second derivative computation nor seperate support or query sets.

Since the original publication, there have been many further developments of the MAML approach. For example, first-order MAML (FOMAML) omitted computation of

second derivatives and thus gained significant computation and speed improvements while maintaining similar performance [27]. Reptile is another method that majorly simplified MAML while achieving nearly identical performance and retaining speed of FO-MAML [97]. Similar to FOMAML, Reptile requires no second derivative computation, but also does not require a train-test data split for each of the tasks, allowing for more natural and simpler implementation. Authors proposed serial and batch versions of the Reptile algorithm depending on how meta-tasks are sampled. In short, serial version of Reptile works by 1) randomly choosing a task from a set of predetermined tasks (referred to as *meta-tasks*), 2) training on that task for multiple gradient steps, and 3) updating model weights towards new parameters (Figure 2.10). The batch version of Reptile follows an identical process for the exception that instead of sequentially choosing meta-tasks, several independent copies of the model are pretrained on all meta-tasks in parallel (further information on Reptile provided in Chapter 3). Authors of Reptile have showed that use of multiple steps of gradient descent for every task allows the algorithm to pick up on the high-order derivatives without the need to explicitly calculate them [97]. Since the original MAML publication, many applications for gradient-based meta-learning methods (e.g. MAML and its derivatives) have been proposed such as few-shot learning [29, 98, 113], federated learning [28, 77, 165], reinforcement learning [2, 156, 159], and large-scale imbalanced classification problems [76, 83, 142].

2.3.3 Meta-learning in Medical Imaging

Within the medical imaging domain, meta-learning has primarily found uses for image segmentation [29, 160], disease classification [105], and anomaly detection [88] problems. Given the scarcity of medical imaging data and difficulty in acquiring ground truth labels [143], one of the most popular uses for meta-learning in this domain is for few-shot learning applications. Few-shot learning [140] is a type of machine learning that aims to train accurate models from small amounts of data [79, 162]. In a study by Farshad et *al*, the authors used Reptile algorithm [97] to leverage existing organ segmentation

datasets to improve segmentation on an unseen organ while using only few samples (or k shots) [29]. Specifically, they redefine meta-task definition as a collection of 2D samples from a 3D volume (CT or MRI) of a single organ that are used for semantic segmentation. After meta-pretraining on several organs, the model was finetuned on a small set of 15 2D slices of an unseen organ. In comparison to supervised and transfer learning methods, the Reptile approach achieved modest performance boost in terms of segmentation metrics. In another study by Singh *et al.*, a similar approach of using gradient-based meta-learning for pathology classification on several skin lesion datasets [127] outperformed traditional transfer learning methods in few-shot learning scenarios and in the presence of significant class imbalance. Moreover, meta-learning has also found uses in domain adaptation where underlying statistics between source and target domains differ [36], a phenomenon often found within medical imaging due to differences in acquisition protocols, patients, and scarcity of data [64, 82, 166]. In one study, Zhang *et al.* proposed slight modification to meta-task selection in Reptile algorithm for domain adaption in organ segmentation, leading to improved generalization performance in colon and liver organ segmentation when compared to original MAML and Reptile methods [160]. Given all of the above, it is important to note that the majority of meta-learning applications in medical imaging treat different datasets as meta-tasks, rather than using different learning objectives (e.g. classification, regression, etc.) for meta-pretraining purposes.

2.4 Summary

This chapter provided the background information necessary to understand the work of this thesis. We reviewed the foundational knowledge of DL, types of models, and their training procedures. Then, we presented the concept of multimodal fusion in DL, different implementation styles, and applications within computer vision and medical imaging, showing the benefits of using multimodal data with DL in a variety of medical tasks. This chapter then introduced the concepts of transfer learning, multi-task learning,

and meta-learning with the idea of models learning shared representation that can be quickly finetuned on a relevant task. Lastly, previous research with applications of meta-learning both in computer vision and medical imaging is discussed, demonstrating the benefits of the approach.

Chapter 3

Meta-learning for Multimodal Fusion

In this chapter, we describe how meta-learning and tabular clinical information is used to improve performance of image-based DL models on clinical tasks. Specifically, this section provides a detailed overview of Reptile meta-learning algorithm and how it is used to provide clinical context and allow CNN models to learn more informative imaging features from MRI. The process and considerations for selection of clinical features used as meta-tasks is then discussed. Lastly, an overview of meta-learning specific hyperparameters and how each controls the meta-pretraining process is provided.

3.1 Meta-learning for Fusion of Clinical and Imaging Information

As mentioned in Chapter 1.4, CNN models sometimes suffer from learning spurious features from imaging data that are not inherently relevant to the actual problem at hand, which can hamper model performance on the desired task [149]. In order to learn more informative imaging features, make use of additional clinical information, and improve model performance, we adapt a modified version of Reptile [97] algorithm as a way of multimodal fusion of 3D MRI imaging and tabular clinical data. The idea is that the features learned by the CNN model from the MRI can be guided with Reptile-based pre-

training to provide more information about the clinical context of the patient through the clinical data. The pretrained CNN model can then be adapted on a desired related task through finetuning on the same MRI data, and result in more accurate model compared to a model trained on imaging data alone without pretraining. The intuition is, for example, that if a CNN model can first learn to classify disease stage and age of the patient from the imaging data (and thus, learning related features), this knowledge can also be useful for predicting clinical test scores of the patient given the assumption that test scores and age are correlated with disease stage. We provide a visualization of Reptile pretraining scheme in Figure 3.1 for easier interpretation.

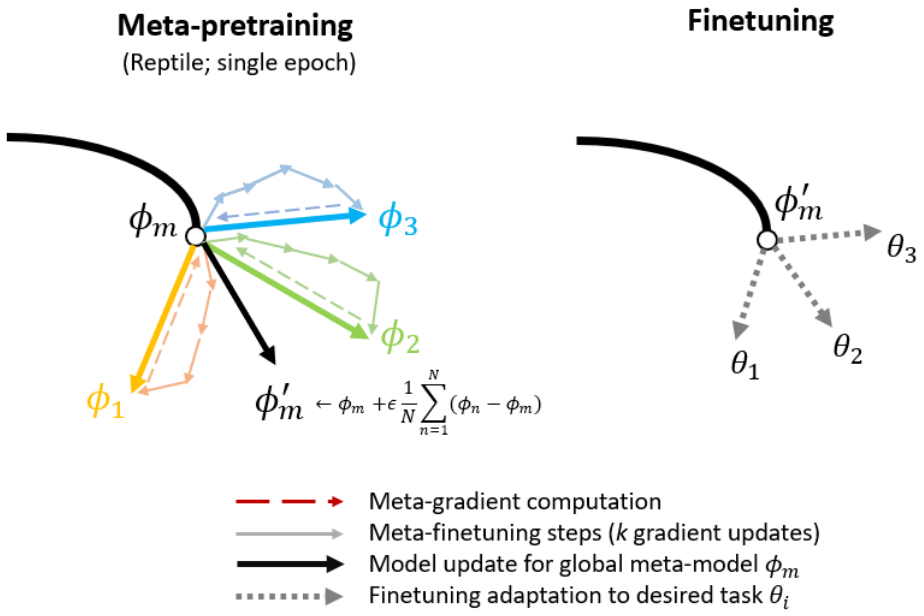

Figure 3.1: Visualization of batch version of Reptile [97] meta-pretraining and finetuning stages. During meta-pretraining, the global meta model ϕ_m is independently trained for k gradient descent operations on prediction each of the meta tasks, resulting in meta-models $\{\phi_1, \phi_2, .., \phi_n\}$. The global meta model ϕ'_m parameters are then updated as an average gradient of all the individual meta models. During finetuning, the pretrained model ϕ_m can be adapted by finetuning on a related desired task (e.g. θ_1)

First introduced in Chapter 2.3.2, Reptile is a gradient-based meta-learning algorithm with the idea of pretraining deep learning models on a distribution of related tasks such

that the model can be quickly adapted on an unseen task [97]. As seen in Figure 3.1, in a single pretraining iteration of batch version of Reptile, the global meta-model ϕ_m (e.g. CNN) is trained in parallel on several meta-tasks (using multiple gradient update steps for each), resulting in unique set of task meta-models $\{\phi_1, \phi_2, .., \phi_n\}$. The global meta-model parameters ϕ_m are then updated with the average gradient across all meta-models, defined as the difference between model parameters of global meta model ϕ_m and task meta-models $\{\phi_1, \phi_2, .., \phi_n\}$. After the pretraining, the model can be adapted to an unseen related target task (e.g. θ_1, θ_2 or θ_3) through finetuning.

In our case, we use batch version of Reptile to pretrain a CNN model (e.g. ResNet [45]) to predict several clinical and demographic features (distribution of tasks) from the MRI data. The set of tasks used for pretraining, referred to as *meta-tasks*, come from the tabular clinical data (e.g. prediction of age, lesion size, test scores), given the assumption that these clinical markers are 1) have a relation to the MRI data, and 2) are related to the desired target task (considerations in selection of clinical data are discussed further in Chapter 3.1.1). Using 3D MRI as input, a ResNet [45] model is pretrained with Reptile such that it generalizes well at across all meta-tasks, that is, predicting all of the clinical features. Lastly, the pretrained ResNet model is finetuned using the same MRI data to adapt it to an unseen desired task (e.g. prediction of diagnosis).

An overview of the proposed method is visualized in Figure 3.2. Steps 1 - 3 show a single epoch of Reptile pretraining of the CNN model on predicting clinical information (meta-tasks). Before pretraining, learnable parameters of a global meta-model ϕ_m (e.g. ResNet CNN) are initialized in a traditional manner (e.g Kaiming [44]), following which we create $\{\phi_1, ..., \phi_n\}_N$ copies (*meta-models*) of the the global meta-model. Using 3D MRI volume as input in step 1, several *meta-models* are trained in parallel (one model ϕ_n per meta-task n for total of N meta-tasks) with a set number of gradient updates. Happening at the end of the pretraining epoch, step 2 performs model parameter update of the *global* meta-model ϕ_m with the exponential moving average (EMA) of all of the individual meta-models $\{\phi_1, ..., \phi_n\}_N$. In step 3, learnable parameters of the meta-models $\{\phi_1, ..., \phi_n\}_N$ are

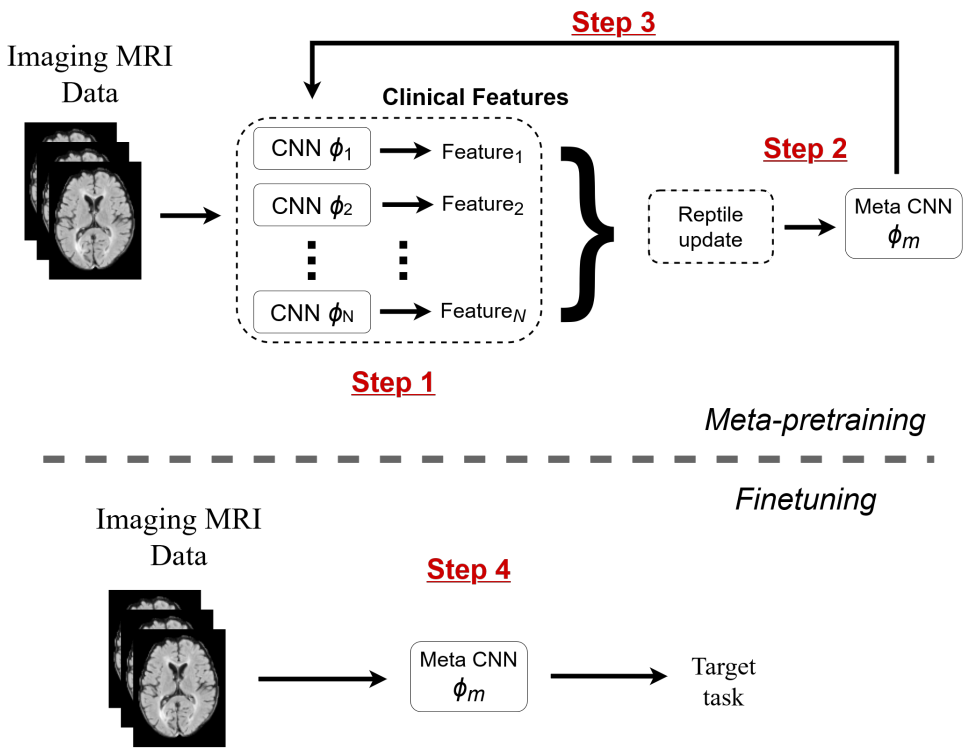

Figure 3.2: Overview of the proposed meta-learning method for fusion of imaging and clinical data using batch version of Reptile; **Top**: meta-pretraining stage to pretrain a ResNet CNN meta-model ϕ_m using MRI as imaging input and tabular clinical information (e.g. age, lesion volume, etc.) as meta-tasks; **Bottom**: pretrained meta-model ϕ_m is finetuned using the same MRI data as imaging input to predict the desired target task (e.g. Gad lesion presence detection).

set to the newly updated parameters of the global meta-model ϕ_m. Pretraining procedure in steps 1 - 3 repeats for a set number of epochs, determined by the user. Finally in step 4, the same MRI data used in pretraining stage is used for finetuning the meta-pretrained model ϕ_m on a desired end task (e.g. lesion detection). Implementation details of batch version of Reptile pretraining are described in detail in Algorithm 1.

We define ϕ_m as the parameters of the global meta-model to be pretrained and used later for finetuning, $\{n_1, n_2, ...\}_N$ is a set of N meta-tasks, U_k operator denotes k-steps of an optimization algorithm (e.g. SGD [4] or Adam [66]), and ϵ is referred to as *meta learning rate*. First, the global meta-model parameters ϕ_m are initialized in a standard manner (e.g. Kaiming initialization [44]). Per one training epoch, global meta-model ϕ_m

Algorithm 1 Batch version of Reptile algorithm [97]

Initialize ϕ_m, initial vector of global meta-model parameters
$\phi_{n=1,2...N} \leftarrow \phi_m$ ▷ Initialize meta-task models parameters to ϕ_m
for epoch=1,2,3.. **do**
  $\phi_{n=1,2...N} \leftarrow \phi_m$ ▷ Reset meta-task model parameters to ϕ_m
  for $n \in N$ **do**
    $\tilde{\phi_n} = U_k^n(\phi_n)$ ▷ Compute $\tilde{\phi_n}$ denoting k training steps of Adam for task n
  end for
  $\phi_m \leftarrow \phi_m + \epsilon \frac{1}{N} \sum_{n=1}^{N} \omega_n(\tilde{\phi_n} - \phi_m)$ ▷ EMA update of ϕ_m meta-model parameters
end for

is trained in parallel for k gradient updates of an optimization algorithm for each task n (represented by the operator U_k^n). The model that has undergone k gradient updates on a meta-task n is referred to as ϕ_n meta-model. In total, there are N parallel meta-models $\{\phi_{n=1}, \phi_{n=2}, ...\}_N$, one per each task n that are trained for k steps. At the end of the training epoch, parameters of the global meta-model ϕ_m are updated in an exponential moving average (EMA) scheme with the uniform weights $\omega_n = 1$ across parameters of task meta-models $\{\phi_{n=1}, \phi_{n=2}, ...\}_N$, as seen in Equation 3.1:

$$\phi_m \leftarrow \phi_m + \epsilon \frac{1}{N} \sum_{n=1}^{N} \omega_n(\tilde{\phi_n} - \phi_m) \tag{3.1}$$

The intuition is to train the global meta-model ϕ_m such that it can generalize and perform well with all meta-tasks, thus, it is updated in the direction of the average gradient of all meta-tasks. The delta between each task meta-model ϕ_n and global meta-model ϕ_m parameters are summed over all the meta-tasks and then normalized by the number of the meta-tasks N. The magnitude of the parameter update of the global meta-model ϕ_m is controlled by *meta learning rate* hyperparameter ϵ ($\epsilon = 1$ meaning ϕ_m is fully set by the average of the task meta-models). Lastly, prior to the beginning of the next training epoch, all task meta-models $\{\phi_{n=1}, \phi_{n=2}, ...\}_N$ parameters are reset to match the updated parameters of global meta-model ϕ_m. This process repeats for a set number of pretraining epochs selected by the user.

3.1.1 Selection of Meta-tasks

As mentioned in Chapter 2.2, one of the important design choices for multimodal fusion is selection of which features/modalities to fuse. In our case, this means the selection of which clinical or demographic features to use with the MRI data. The question then becomes as to which markers are 1) related to, and 2) potentially beneficial when used alongside imaging data. Recall meta-pretraining stage in Figure 3.1; the global meta-model ϕ_m aims to learn useful representation that is shared between all meta-tasks ϕ_1, ϕ_2, ϕ_3. Given the original assumption that different modalities (in this case, clinical features used as meta-tasks) provide useful complimentary information, a pretrained model should in theory obtain better performance when fine tuned on a *related* target task. However, if one or all of the meta-tasks are selected incorrectly (e.g. have no relation to other tasks or the input data), this can instead "confuse" the model and be detrimental to performance on the target task. This is known as *negative transfer* [161] and is common with transfer learning methods [167].

Given the assumption that imaging data carries useful information for the target task, the clinical meta-tasks should also be directly related to the input imaging data. For example, if the target task is to predict treatment response of a patient from a baseline time point MRI scan (when no treatment effect was yet visible), treatment information cannot be provided as a meta-task since the imaging data (baseline time point MRI scan) has no relation to the treatment information. However, treatment information could be used as a meta-task if the input imaging data also included MRI scans sometime after treatment effect appears (e.g. seen as decrease in the number of new or enlarged T2 (NE-T2) lesions in the MRI [118]). Lastly, since our approach aims to enhance the imaging features by using clinical data, the imaging data is assumed to be the main (or at least partial) driver behind model performance on the target task (that is, imaging data has correlations with the target task).

Compared to traditional multimodal fusion methods described in Chapter 2.2, correct selection of relevant clinical features carries more importance in the meta-learning

scheme. This is due to the fact that the pretraining dynamics in a meta-learning approach are not directly impacted by the performance of the model on the desired end task, since there are two separate training stages. In comparison, multimodal methods seen in Chapter 2.2 use the gradient of the loss of the target task as a direct feedback signal to adjust the impact of individual modalities automatically (e.g. early or joint fusion). With Reptile, one way to control the impact of individual clinical features is by adjusting meta-task weights ω_n during EMA update stage in Algorithm 1, assigning higher weight to tasks of more importance (e.g. higher relevance to the end task). Selecting ω_n weights is done manually as part of hyperparameter optimization. For example, original Reptile implementation assigns uniform ω_n across all meta-tasks, meaning each meta-task has identical contribution to the global meta-model. In contrast, one approach presented by Farshad *et al.* uses inverse distance weighting (IDW) scheme to control meta-task importance by giving more weight to meta-models with parameters closer to the parameters of the global meta-model [29], measured by the squared difference between the weights:

$$\omega_n = \frac{1}{\sum_{i=1}^{I}(\phi_{n,i} - \phi_{m,i})^2} \tag{3.2}$$

In Equation 3.2 $\phi_{m,i}$ is the *i*-th weight of the global meta-model ϕ_m, and $\phi_{n,i}$ is the *i*-th weight of the *n*-th task meta-model. The meta-task weights ω_n for n-th task are then normalized across all N meta-tasks by dividing each ω_n by the sum of all other meta-weights $\sum_{n=1}^{N} \omega_n$ [29]. Our exhaustive attempts to use similar task weighting scheme lead to highly unstable and failed training, and it was decided to use the average meta-task weighting scheme as was done in the original Reptile implementation [97] to reduce the scope of our study and simplify hyperparameter tuning.

3.1.2 Controlling Meta-pretraining

As mentioned in Chapter 2.1.2, the process of training a model in a traditional scheme is largely controlled by the number of epochs, learning rate, choice of optimizer, and

regularization (e.g. L2 weight decay) hyperparameters. With meta-learning, there are also additional hyperparameters that affect the learning process during meta-pretraining: number of gradient updates *k*, meta learning rate ϵ, and the number of meta-pretraining epochs. Number of gradient updates per task meta-model, k, in our case is controlled by the amount of training data provided and the batch size (given same amount of data, larger batch size means less gradient updates but more stable training [72]). Meta learning rate, ϵ, controls the amount of change model parameters of ϕ_m receive from the trained meta-models $\{\phi_{n=1}, \phi_{n=2}, ...\}_N$ in one epoch, identical to how learning rate controls the speed of learning in traditional deep learning training. Lastly, the number of meta-pretraining epochs controls duration meta-pretraining process and encompasses the impact of all of the previously mentioned hyperparameters. As with other hyperparameters, selecting correct values for the above is part of the overall hyperparameter tuning process.

3.2 Summary

This chapter provided detailed description of our proposed method of adapting Reptile algorithm for improving performance of image-based DL models for interpretation of medical imaging tasks. It outlined the process and considerations of selecting meta-tasks, as well as hyperparameters controlling the pretraining process. In this thesis, relevant clinical markers and demographic information are used as meta-tasks during meta-pretraining of a CNN model. By using clinical features for model pretraining, CNN is guided to learn more informative imaging features with regards to the target clinical task (e.g. lesion detection). The pretrained model can then be finetuned on a relevant target task and achieve better performance compared to models trained only using the imaging data. The clinical features must be chosen carefully in order to adhere to the assumption that they will provide useful complimentary information to the desired target task, as well as being directly related to the content of the imaging data. Lastly, amount of

meta-pretraining can be controlled through three main meta-hyperparameters that can be selected as part of a wider hyperparameter tuning process.

Chapter 4

Implementation and Experimental Details

This chapter provides the implementation and experimentation details for the proposed meta-learning method in Chapter 3. Through experimentation, we hope to show that the proposed meta-learning approach can perform better than models trained with imaging-only data. By using MS and AD datasets described in this chapter, we evaluate performance by training all methods on three example target tasks: 1) detection of Gad lesion presence in MS patients, 2) prediction of future lesion activity in MS patients, and 3) regression of ADAS-13 and MMSE cognitive scores in AD patients. Given that our method uses multimodal data, we also compare performance of the proposed method against existing deep learning methods that make use of medical imaging and clinical data together. As a surrogate measure to understand how informative medical imaging and tabular clinical data are independently of each other, we first trained separate models using either imaging-only data or clinical-only data (referred to as *unimodal* methods). For methods that use both modalities (imaging and clinical data) as model input, we trained examples of early, joint, and late fusion multimodal methods described in Section 2.2. Furthermore, given that the proposed meta-learning method is an example of a transfer learning approach, we also compare its performance to existing transfer learning techniques such

as multi-task learning and multi-task pretraining. All experiments are implemented in Python using PyTorch v1.10 [100] and MONAI v0.9.1 [17] frameworks.

4.1 Multiple Sclerosis Dataset

We make use of five proprietary MS clinical trial datasets that have been pooled together, resulting in a total of 3560 patients (listed in Table 4.1). Each clinical trial contains patient samples with longitudinal data of five MRI sequences: T2-weighted (T2-w), T1-weighted (T1-w), T1 with Gadolinium contrast agent (T1-ce), PD, and post-contrast FLAIR. For experiments on detection of Gad lesion presence, we use all patients across all treatment arms listed in Table 4.2, meanwhile experiments on prediction of future NE-T2 lesion activity only make use of placebo (no treatment) patients as to avoid taking into account treatment effects. Nonetheless, all MS experiments make use of baseline time point MRI scans as the imaging data. In addition to MRI, every patient sample also contains tabulated longitudinal clinical data, containing information such as demographics (e.g. age, sex), clinical test results, disease stage, and MRI-derived features (discussed later in more detail in Section 4.4). The above MRI sequences were originally obtained at 1mm $\times$ 1mm $\times$ 1mm resolution, and were then down scaled to 2mm $\times$ 2mm $\times$ 2mm resolution. Prior to their use, the MRI scans were preprocessed in a consistent manner to minimize acquisition effects due to differences in scanners, given that each trial merged data from dozens of different study sites and MRI scanners. For all scans, the steps were as follows: denoising [85], intensity heterogeneity correction [128], and intensity normalization into [0,100] range. Next, FLAIR, PD, and T2-w scans were co-registered to T1-w scan using a six-parameter rigid registration [16], after which the T1-w scans were registered to an average template stereotaxic space [19]. Lastly, all scans were then resampled onto a 1mm isotropic grid. For our experiments, we also z-score standardized all MRI volumes by subtracting mean and dividing by standard deviation of the pixel intensity values within

the brain region, followed by skull stripping with the brain mask to leave only the brain structures in the image.

Table 4.1: Trial names, MS disease phenotypes, number of patients, and number of unique study sites per trial in the MS dataset.

Trial Name	Disease Phenotype	Patients	Unique Sites
BRAVO	Relapsing Remitting	992	150
DEFINE_ENDORSE		408	77
ADVANCE_ATTAIN		1266	183
ASCEND	Secondary Progressive	599	154
ORATORIO	Primary Progressive	295	16

4.2 Alzheimer's Disease Dataset

For the task of regression of cognitive scores, we use a subset of public Alzheimer's Disease (AD) dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [104] database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

The AD dataset is made up of 722 patient samples containing baseline time point T1-w MRI volumes obtained at 1mm $\times$ 1mm $\times$ 1mm resolution which were then downscaled to 2mm $\times$ 2mm $\times$ 2mm. Without affecting visibility of vital structures in the MRI (e.g. hippocampus), reduced dimensionality of the input MRI allowed for reduced number of model parameters, leading to faster training time and experimentation process. All MRI volumes were brain-extracted and linearly registered to MNI152 [84] space. For our experiments, MRI volumes are also z-score standardized within the brain region and then skull stripped by using a brain mask. Out of 766 patients, 197 are diagnosed as cognitively

normal (CN), 409 with mild cognitive impairment (MCI), and 116 with Alzheimer's disease (AD). The dataset also contains clinical records of the patient in tabular form, taken at the time of MRI acquisition. This includes imaging summary statistics (e.g. hippocampus volume), laboratory results, genetic markers, demographics information, and clinical test scores (discussed further in Section 4.4).

4.3 Selection of Target Tasks

To test our proposed method, we selected three medical imaging interpretation tasks that we hypothesize could benefit from inclusion of clinical context information (as originally introduced in Chapters 1.2 and 1.3): 1) detection of presence of Gad brain lesions in MS patients, 2) prediction of future (48 weeks later) new and enlarging T2 (NE-T2) lesion activity in MS patients, and 3) estimation of disease severity scores Alzheimer's Disease Assessment Scale (ADAS-13) [138] and Mini-Mental State Examination (MMSE) [32] in AD patients, all from baseline MRI scans.

For task #1, previous studies using DL methods generally relied on lesion segmentation [35, 60, 61] (achieved 0.93 recall) and lesion counting [92] (achieved 0.86 recall and 86.3 F1 score) for detection of Gad lesions. In our case, we simplify task #1 by designing it as a binary classification problem to detect from the MRI data whether the patient has real Gad lesions present. The ground truth label is the binarized count of Gad lesions, with patients that have one or more Gad lesions belonging to the positive class and negative class otherwise. We selected this threshold to allow as many positive class cases as possible in order to tackle severe class imbalance (seen in Table 4.2). This task was selected because we hypothesize that inclusion of clinical context can help discern uncertainty in identifying whether hyperintensity areas in MRI are real Gad lesions or not (described in more detail in Section 1.2 and visualized in Figure 1.1). Given that presence of Gad lesions is one of the indicators of MS disease worsening, detection of their presence can

alert the treatment providers to take a closer look at the MRI data and adjust treatment if necessary.

Similarly to task #1, prediction of future NE-T2 lesion activity has primarily been attempted through lesion segmentation [39, 119, 122], lesion count regression [23, 25], and lesion presence classification [123] from multi-sequence MRI data. Out of the listed studies, the one [25] with the closest equivalent experiment task (prediction of future NE-T2 of placebo patients) achieved ROC AUC of 0.836, meanwhile the next closest study [123] attained 80.21% accuracy. For our purposes, task #2 was designed as a binary classification problem to predict whether a patient will have future NE-T2 lesion activity. The ground truth label is a binarized NE-T2 lesion count at 48 weeks after the initial MRI scan was taken. According to the MS guidelines [33], a cutoff of three or more NE-T2 lesions is used as an indicator of minimal evidence of disease activity, and as such, patients with NE-T2 count ≥ 3 are considered as positive class (*active* patient) and negative class otherwise (an example of an active patient MRI with NE-T2 lesions can be seen in Figure 1.2). We selected this task as we hypothesize that prediction of future NE-T2 lesion activity can benefit from knowledge of clinical history (e.g. disease stage, age) that can be predictive of disease activity in addition to MRI information. With the ability to predict whether the patient will have NE-T2 lesion activity in the future, this information can be of use to physicians in order to help select appropriate treatment earlier.

For task #3, we focused on estimating ADAS-13 and MMSE scores given that in medical practice, patient symptoms are more likely treated based on clinical assessments rather than on a specific diagnosis [133]. MMSE is a commonly used cognitive assessment for diagnosis of Alzheimer's disease with the scores ranging from 0 to 30, with lower score indicating more severe impairment. ADAS-13 is version of ADAS-cog test used for assessment of severity of dementia symptoms, with scores ranging from 0 to 85 and higher score indicating greater cognitive impairment. Previous studies [24, 73, 90, 136] generally focused on prediction of future MMSE and ADAS scores (mainly MMSE) by using multi-task and joint learning methods from MRI and clinical data. By learning diagnosis clas-

sification in addition to cognitive score prediction, [24] achieved 3.27 RMSE for MMSE prediction, meanwhile prediction of several cognitive scores at once by [73] obtained MMSE MAE of 1.92. As closer equivalents to our target task of estimating MMSE and ADAS-13, [136] obtained 2.50 RMSE when estimating MMSE in conjunction with disease classification task, and [90] achieved 2.15 and 7.34 RMSE in joint MMSE and ADAS-13 regression respectively. Similarly to other target tasks, we believe that estimating ADAS-13 and MMSE scores from MRI data can benefit from clinical context due to uncertainties present in imaging information. For example, reduced hippocampal volume is a known biomarker for AD [34] and is determined by examining the size of hippocampus visible in the MRI. Given that hippocampal volume also decreases naturally as humans age older [34], this creates ambiguity when trying to diagnose the patient using MRI data alone. Knowledge of clinical context such as patients age can greatly help confirm the diagnosis since visible decrease in hippocampal volume of a 20 year old patient is much more indicative of AD diagnosis compared to the same decrease in a 70 year old patient. Comparison between MRI of cognitively normal patient and a patient with Alzheimer's disease is shown in Figure 1.3.

4.4 Selection of Supporting Clinical Data

As first mentioned in Chapters 2.2 and 3.1.1, selection of supporting clinical data to be used as model input is done on the assumption that it is related to the target task and ideally, also the other input modality (in our case, MRI scans). Given the motivation of our method to guide CNN model in learning more informative features from MRI, we also make use of image-derived data found in the MS and AD datasets in order to investigate effects between non-imaging (e.g. age, sex, disease diagnosis) and image-derived clinical data (e.g. T2 lesion volume, hippocampus volume) on model performance. Note that throughout this thesis, we often refer to supporting clinical data (e.g. age, lesion volume, cognitive scores) as clinical *features* or clinical *markers*.

4.4.1 Multiple Sclerosis

For prediction of future NE-T2 lesion activity (target task #2), we selected Expanded Disability Status Scale (EDSS) score [69], T2 lesion volume, age, MS disease stage (phenotype, e.g. RRMS), and Gad lesion count at baseline time point as support clinical data, given that majority of these have been successfully used in conjunction with MRI data for prediction of MS treatment efficacy in previous studies [25]. These tabular clinical features provide valuable clinical context to the imaging model as follows. EDSS value provides a score quantifying levels of disability and by extension the severity of MS [69]. Volume of T2 lesions has been found to be a robust marker of MS progression [38], and thus, directly related to MS lesion activity. The disease phenotype (e.g. RRMS, SPMS) itself is in part determined by the lesion load of the patient. Age of the patient has been shown to be one of the prognostic factors for MS [5, 43]. Lastly, Gad lesion count can reflect the level of lesional MRI activity (how many new lesions are forming) and can indicate the stage of MS disease [30]. For detection of Gad lesion presence (target task #1), we used the same supporting clinical data as described above for the same reasons with one exception. Gad lesion count was excluded as one of the supporting clinical features due to the fact that the target task #1 aims to predict presence of Gad lesions, which is in itself a binarized version of Gad lesion count. All of the above clinical and demographic data were measured at the baseline time point (the initial visit for acquisition of MRI), with the exception of future NE-T2 lesion count, which was recorded at the next consecutive time point 48 weeks later. The relevant statistics are presented in Table 4.2.

4.4.2 Alzheimer's Disease

Supporting clinical information for ADAS-13 and MMSE regression are divided into two subsets: clinical and image-derived features. For clinical features, we selected age, years of education, sex, and presence of apolipoprotein E4 (APOE4) gene variant [65] as these are known risk factors and markers for development of AD and dementia [6]. For imaging

Table 4.2: Selected tabular clinical data and demographics statistics for MS dataset

Feature	Statistics
EDSS	Mean: 3.24, Std. Dev.: 1.71
Age (years)	Mean: 39.38, Std. Dev.: 10.01
T2 lesion volume (ml^3)	Mean: 11.00, Std. Dev.: 12.79
MS Phenotype	Number of patients: **PPMS**: 294; **RRMS**: 2666; **SPMS**: 599
Treatment	Number of patients per treatment: **Dimethyl Fumarate:** 265 **Interferon Beta-1a:** 338 **Laquinimod:**317 **Natalizumab:** 308 **Ocrelizumab:** 207 **Peginterferon Beta-1a:** 1263 **Placebo:** 862
Gad-enhanced lesions	Number of patients with/out Gad-enhanced lesions: **Present**: 1223 (34.4%); **Absent**: 2337
Future lesion activity	Number of patients with/out future NE-T2 lesion activity: **Active**: 1070 (30.5%); **Inactive**: 2490

summary data, we used measurements of hippocampus, ventricles, entorhinal cortex, and whole brain volumes since volume measurement of select brain structures is one way to quantify brain atrophy, and is one of the methods to diagnose stages of AD [56]. Relevant statistics of the above data are presented in Table 4.3.

Table 4.3: Selected tabular clinical data and demographics statistics for AD dataset

Feature	Statistics
Hippocampus volume (ml^3)	Mean: 6795.31, Std. Dev.: 1158.58
Ventricle volume (ml^3)	Mean: 38198.59, Std. Dev.: 21229.54
Whole brain volume (ml^3)	Mean: 1024495.88, Std. Dev.: 109887.25
Entorhinal cortex volume (ml^3)	Mean: 3468.82, Std. Dev: 757.107
Age (years)	Mean: 72.99, Std. Dev: 7.04
Education (years)	Mean: 15.95, Std. Dev: 2.80
Sex	**Male**: 337, **Female**: 385
ADAS-13	Mean: 16.41, Std. Dev: 9.284
MMSE	Mean: 27.31, Std. Dev: 2.61
APOE4 presence	Number of patients: **Absent**: 391; **Present**: 331;

4.5 Model Architecture

As the base CNN model, we use a popular ResNet [45] architecture as seen in Figure 4.1. It consists of two major components: CNN backbone for extraction of MRI features and an MLP classification head. The CNN backbone is made up of four residual blocks (bottom left of Figure 4.1), along with adaptive average pooling and flattening operations at the end. The MLP head consists of two fully-connected layers with leaky ReLU activation in the middle. The model hyperparameters are outlined in the Table 4.4 below.

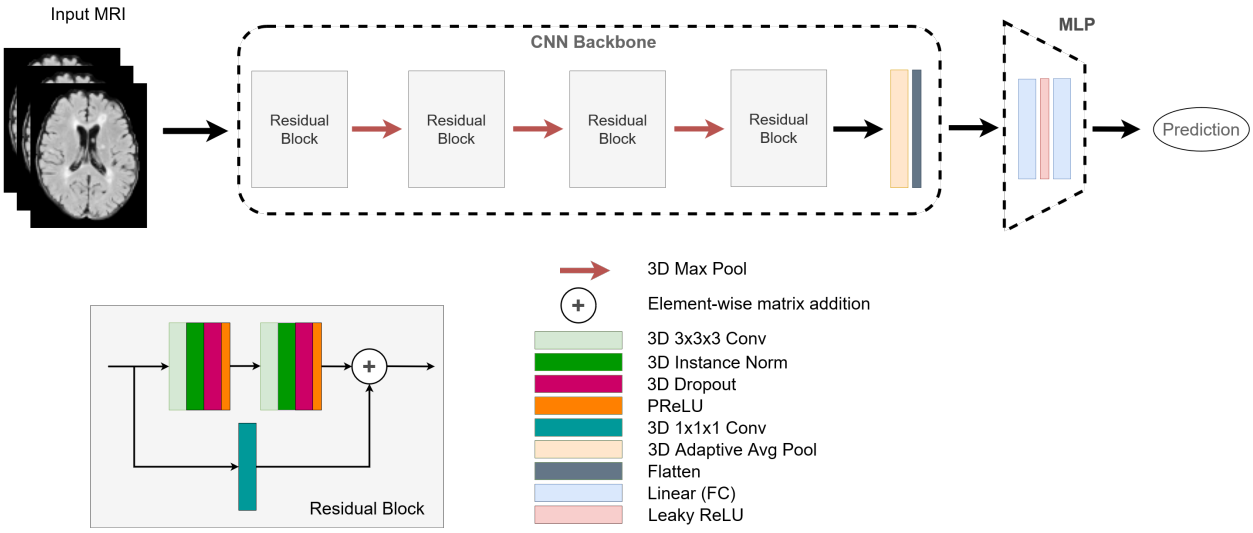

Figure 4.1: Diagram of the CNN architecture. **Top:** ResNet model with MRI-only input; minor modifications to the MLP classification head are made for different experiments. **Bottom left:** architecture of the residual block.

Table 4.4: Model architecture details. Note that for the linear layers in the MLP, the number of input or output nodes changes with respect to the task at hand (e.g. single output node for binary classification, but three for 3-class classification).

Block	Layer	Details
Residual	3x3x3 Conv3d	Channels (block 1 to 4): [32, 64, 128, 256] Stride: 1; Padding: 1
	Dropout	P = 30%
	1x1x1 Conv3d	Same as 3x3x3 Conv3d
MLP	Linear	Layer 1: 256 to 32; Layer 2: 32 to 1*
	Leaky ReLU	Negative slope: 0.01

4.6 Baseline Methods

As a surrogate for quantifying how informative each modality is with regards to the target tasks, we trained *unimodal* models that use only imaging (referred to as *MRI-only*) or only tabular (referred to as *clinical-only*) data as input. For *MRI-only* experiments, the entire model in Figure 4.1 is used, meanwhile, *clinical-only* experiments utilize a 2-layer MLP model (seen in Figure 4.1, without the CNN backbone), both of which are commonly used architectures in for processing imaging and tabular-style data respectively.

Given that our proposed method makes use of multimodal data (imaging and clinical), we compare our method to other existing methods of multimodal fusion in medical imaging by training examples of *early-fusion*, *joint-fusion*, and *late-fusion* models as originally described in Chapter 2.2. For *early fusion* (Figure 4.2**a**), imaging features from the ResNet CNN are concatenated with preprocessed clinical data which are then used as input to the MLP. For *joint fusion*, clinical features are first extracted by an MLP and are then concatenated with ResNet-extracted imaging features prior to being used as input to another MLP (Figure 4.2**b**). For *late fusion* in Figure 4.2**c**, there are two separate models for each modality (ResNet for MRI, MLP for clinical data) both of which predict the target task (for example in classification tasks, these are pre-softmax values). The actual output is then the mean value of the individual model outputs. The process is mirrored for AD experiment for the exception that there are two parallel MLP classification heads (a multi-task approach), one for regression of each of the ADAS-13 and MMSE cognitive scores.

Since the proposed meta-learning approach is a subtype of transfer learning, we also compare it to existing transfer learning methods in the medical imaging domain. One popular approach to learning a shared imaging representation is multi-task learning (described in Chapter 2.3.1). Shown in Figure 4.2**d**, both target tasks and supporting clinical features are predicted simultaneously via parallel MLP classification heads (one per task) and using a shared ResNet backbone. Furthermore, to mirror the pretraining-finetuning

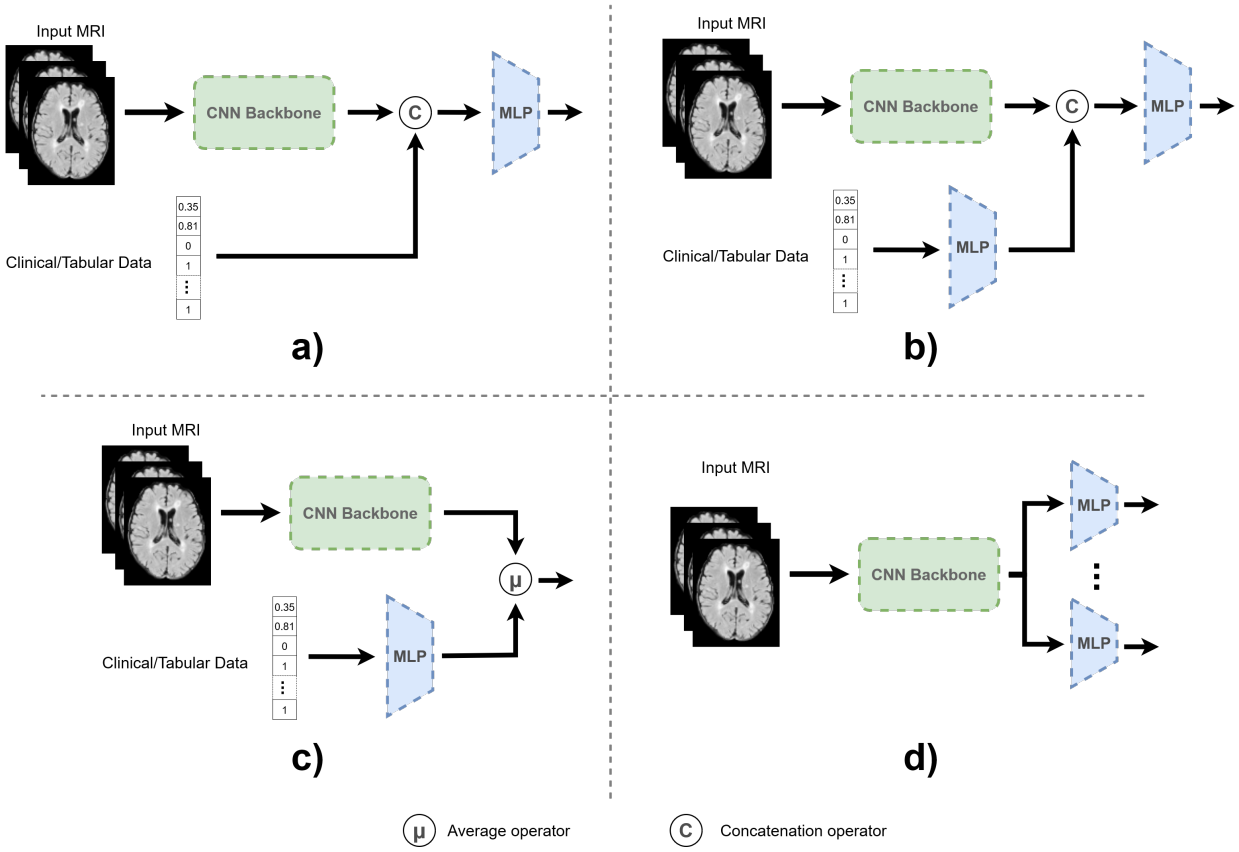

Figure 4.2: Architectures of different multimodal and transfer learning baseline models (**C** - concatenation, μ - mean operator). CNN backbone from Figure 4.1 is used to learn imaging features from MRI; 2-layer MLP blocks are used for clinical feature extraction or feature mixing; **a)** is an example of early fusion by concatenating raw clinical feature vector with imaging features, **b)** is joint fusion by first passing raw clinical data through an MLP to extract features prior to concatenation, **c)** is late fusion where logit predictions are averaged between CNN and clinical models prior to output; **d**) shows architecture of multitask learning approach where both target task and clinical features are predicted at the same time. Details of CNN backbone and MLP networks are shown in Figure 4.1.

process of our proposed method, a similar multi-task baseline method (referred to as *multitask pretrained*) is used. A multi-task model (Figure 4.2**d**) is first pretrained to simultaneously predict all of the supporting clinical features. Afterwards, the multihead MLP blocks are replaced with a newly-initialized MLP block, and the entire model is then fine-

tuned on the target task. With this method, we compare the differences between learning imaging features through a multi-task approach against using a meta-learning approach.

It is important to note that architecture details change slightly depending on the target task. For example, the number of nodes in the output layer can change when used for multiclass classification tasks (e.g. MS phenotype) and single-value regression (e.g. age). However, all of the methods are made up of identical building blocks (e.g. CNN backbone, MLP blocks) as to maintain that the number of trainable model parameters (and by extension, their representation power [9]) of the models remain similar in order to allow for fair comparison between the methods.

4.7 Training Procedures

To ensure evaluation validity and minimize the possibility of distribution shifts, data splits (divided into a 70/10/20% training/validation/testing ratio) were created in a way such that class distributions remained identical between the splits. In the case of task #3 (ADAS-13 and MMSE score regression), we maintained the distribution of patients through disease diagnosis (CN, MCI, and AD) and confirmed that summary statistics of ADAS-13 and MMSE remain similar between data splits. Furthermore, the size of the MS dataset for task #2 (prediction of future NE-T2 lesion activity) was reduced from 3560 patients to only 862 by using only placebo patients in order to avoid introducing more task complexity due to presence of treatment. Note that for all MS experiments, we only used post-contrast FLAIR MRI scans in order to reduce the overall dimensionality and complexity of the imaging input compared to using all available sequences. Given the fact that the FLAIR scans were acquired after the patient was administered Gadolinium contrast agent (hence *post-contrast*), these scans still capture the relevant information about both Gad-enhanced and T2 lesions in the scan (both seen as hyperintensities in the image).

During the model training stage, image augmentations are applied to artificially increase our dataset size and improve robustness to image perturbations [126]. We make

use of random flipping along saggital axis (left and right hemispheres) with the probability of 30%, and apply random Gaussian noise with $\mu = 0$ and $\sigma = 0.1$ with 30% probability. Any numerical data (age, EDSS score, years of education, brain volumes, cognitive test scores) are z-score standardized with their respective mean and standard deviation statistics (seen in Tables 4.2 and 4.3), meanwhile categorical data (MS phenotype, indicator of APOe4 gene presence, sex) are converted to a one-hot encoded representation.

For all experiments, model parameters were first initialized using Kaiming [44] initialization with an identical random seed prior to training. For methods trained directly on the target task (no pre-training), we used the following hyperparameters for model training (with minor modifications due to hyperparameter tuning): Adam optimizer [66], learning rate $\alpha = 3 \times 10^{-4}$, L2 weight decay $\gamma = 2 \times 10^{-5}$, step learning rate decay rate of 0.995 after every epoch, and a minibatch size $b = 6$ (this controls the number of gradient updates per meta-epoch given a constant dataset size). While the models were trained for 800 epochs in order to observe full training dynamics, evaluation metrics were taken when the model achieved lowest validation loss. The same hyperparameters were used for model pretraining with the exception of learning rate $\alpha = 1 \times 10^{-3}$ and L2 weight decay $\gamma = 2 \times 10^{-4}$, in addition to meta learning rate $\epsilon = 0.75$ and 75 epochs. For finetuning, a similar setup is used but instead with learning rate $\alpha = 2 \times 10^{-4}$ and L2 weight decay $\gamma = 3 \times 10^{-5}$ to tune all the trainable parameters. Note that the same dataset is used for both pretraining and finetuning stages.

Depending on the task, we selected the appropriate loss function for the training process. Binary cross entropy (BCE) is used for binary classification tasks (e.g. detection of Gad lesion presence), mean squared error (MSE) for regression (e.g. EDSS score), and cross entropy (CE) loss for multi-class classification (e.g. MS phenotype classification). Specifically for target task of joint ADAS-13 and MMSE scores regression, we utilized root mean squared error (RMSE) to mirror what has already been used in previous studies for this exact problem [24, 90]. For the *multitask* and *multitask pretrained* methods, the total loss used for backpropagation is the sum of losses from all of the tasks as seen in

Equation 4.1, where N is the set of learned tasks (e.g. prediction of multiple supporting clinical markers).

$$\mathcal{L}_{total} = \sum_{n=1}^{N} \mathcal{L}_n \tag{4.1}$$

4.8 Performance Evaluation

Given that both of the target MS tasks are binary classification tasks, we measure performance with metrics designed for binary classification problems (introduced in Chapter 2.1.3). Both of the target MS tasks experience severe class imbalance where the positive class (e.g. patient with Gad lesions detected) is much less prevalent than the negative class (see Table 4.2 for details). For both of the MS tasks, we put the emphasis on correct detection of the positive class because correct detection of Gad lesion presence or potential future NE-T2 activity in a patient can be a sign for additional treatment intervention by the medical staff. As such, we select metrics that focus more on the accuracy of prediction for the positive class, which are precision-recall (PR) and F1 score metrics. In contrast, we also use ROC metric as it is not biased towards minority nor majority class, and is one of the standard metrics used for binary classification. With both PR and ROC, we make use of AUC in order to eliminate the effects of binarization threshold selection. It is important to note that binarization thresholds can vary from metric to metric when selecting for the highest metric value. For regression tasks, performance is evaluated with either MSE or RMSE depending on the task. Detailed description of the above metrics is provided in the Chapter 2.1.3. Unless stated otherwise, all models are trained in a 4-fold cross validation manner using their respective MS or AD datasets (data folds across experiments are identical).

4.9 Summary

This chapter presented implementation details for experiments in the following chapters, and discussed the datasets used, data preprocessing steps, description of baseline methods, training procedures, as well as evaluation metrics. Information provided in this chapter is of major importance to other researches attempting to replicate results in this thesis. In the following chapter, the models described in this chapter are used to compare performance and feasibility of the proposed meta-learning approach against existing unimodal, multimodal, and transfer learning methods on the target tasks, compare robustness in varying data regimes, and investigate the importance of individual clinical markers for the target task performance.

Chapter 5

Experimental Results and Discussion

In this chapter, we present and discuss results of experiments and methods discussed in Chapter 4. We first show performance metrics of the proposed meta-learning approach along with unimodal, multimodal, and transfer learning methods on the target tasks of detection of Gad lesion presence, prediction of future NE-T2 lesion activity, and regression of ADAS-13 and MMSE cognitive scores, with the belief that meta-learning can perform better than unimodal and potentially transfer learning methods. This chapter then investigates the effects of individual clinical features and the selection of supporting clinical data on the target task performance, where we hope to gain more insight on how to best select the supporting clinical data for the meta-learning approach. Following is a short exploratory studying to investigate how the meta-learning approach performs across varying data regimes in comparison to the other baseline methods. Lastly, this chapter concludes with discussion about limitations of meta-learning approach and considerations about its use.

5.1 Comparison of Multimodal Fusion Methods

This section presents experimental results of the proposed meta-learning method to evaluate its performance on the tasks of 1) detection of Gad lesion presence, 2) prediction

of future NE-T2 lesion activity, and 3) regression of ADAS-13 and MMSE cognitive test scores. Results from MRI-only or clinical-only methods are used as surrogate measurements of how informative each modality is. We hypothesize that the proposed meta-learning approach can perform better than the MRI-only method, as well as achieve competitive performance with other transfer learning methods (multitask and multitask pretrained). Furthermore, we also compare performance of the above methods to examples of early, joint, and late multimodal fusion approaches. Descriptions of different methods are presented in Chapter 4. Performance on binary classification tasks (tasks #1 and #2) is evaluated through ROC AUC, PR AUC, and F1 score metrics. Performance on regression task # 3 is evaluated using RMSE metric on each individual cognitive score.

5.1.1 Results and Discussion

Detection of Gadolinium-enhanced Lesion Presence

Results on the task of detection of Gad lesion presence are shown in Table 5.1 as mean and standard deviation across all folds. All experiments use baseline post-contrast FLAIR MRI and clinical data recorded at the time of MRI acquisition.

Table 5.1: Performance metrics of various unimodal, multimodal, and transfer learning techniques for detection of Gad brain lesion presence. Performance is measured using 4-fold cross validation showing mean and standard deviation respectively. Gray are unimodal methods, cyan are multimodal methods, and green are transfer learning methods. Best results in bold, arrow direction indicates that higher value is better.

Method	**ROC AUC** ↑	**PR AUC** ↑	**F1** ↑
MRI-only	0.822±0.006	0.763±0.008	0.673±0.004
Clinical-only	0.777±0.023	0.659±0.026	0.651±0.022
Early Fusion	0.829±0.004	0.780±0.005	0.689±0.007
Joint Fusion	0.838±0.002	0.784±0.007	0.682±0.005
Late Fusion	0.833±0.005	0.778±0.001	0.684±0.004
Multitask	0.748±0.011	0.677±0.010	**0.697±0.009**
Multitask Pretrained	0.758±0.012	0.671±0.015	0.649±0.008
Meta-learning	**0.848±0.007**	**0.797±0.005**	0.694±0.004

Comparing the methods using only one modality, MRI-only approach performed significantly better than clinical-only approach across all metrics. Given that Gad lesions are visible hyperintensities on the MRI (see Figure 1.1), it is logical that a model using MRI data as input performs better as there is a direct relationship between the input and the ground truth label (presence of Gad lesions), in comparison to learning the indirect relationships from the clinical data. Furthermore, clinical-only method showed the highest variance out of all other methods, which can be a sign of overfitting on the training dataset and bad generalization to new data.

In general, multimodal non-transfer learning methods (early, joint, and late fusion) have performed 1 - 2% better than the single modality methods across all metrics. While its a marginal improvement, this shows the benefits of multimodal methods and using imaging and clinical data together. Interestingly, all of the above multimodal methods performed very similarly to each other with no major benefit to either one on this specific target task.

Comparing the transfer learning methods (multitask, multitask pretrained, and meta-learning), multitask methods performed the worst across all methods (for the exception of F1 score). Given that both multitask methods only use MRI as input, the significant drop in performance compared to MRI-only approach is a potential sign of negative transfer [161] phenomenon (introduced in Chapter 2.3), where the shared imaging representation learned by the multitask approach was in-fact detrimental to the detection of Gad lesion presence task. In contrast, the proposed meta-learning approach performed the best in two out of three metrics, marginally outperforming the next best method by 1 - 2% in ROC AUC and PR AUC and performing second-best in F1 score. While it is unknown whether the above findings are statistically significant, the improved results of meta-learning method compared to MRI-only hint that pretraining the CNN on the supporting clinical data is beneficial to learning the target task. However, observing worse performance metrics of the other transfer learning methods (multitask and multitask pretrained) when compared to meta-learning, we believe that the differences in the pretrain-

ing mechanism (e.g. multitask or meta-learning approach) are one of the determining factors whether pretraining the model will be beneficial or not to the target task.

Prediction of Future NE-T2 Lesion Activity

Presented in Table 5.2 are the results of meta-learning and various baseline methods on prediction of future (48 weeks later) NE-T2 lesion activity from baseline MRI and clinical data. As described in detail in Chapter 4.4, we selected T2 lesion volume, EDSS score, age, MS phenotype, and Gad lesion count as supporting clinical data. Compared to experiment in Section 5.1.1, we only utilized placebo patients (no treatment) for this experiment in order to simplify the task by removing treatment effects, resulting in 862 samples with 332 considered as active ($\geq$ 3 NE-T2 lesions in the future).

Table 5.2: Performance metrics of unimodal, multimodal, and transfer learning techniques for prediction of future NE-T2 lesion activity on placebo patients. Performance is measured using 4-fold cross validation showing mean and standard deviation respectively. Gray are unimodal methods, cyan are multimodal methods, and green are transfer learning methods. Best results in bold, arrow direction indicates that higher value is better.

Method	ROC AUC ↑	PR AUC ↑	F1 ↑
MRI-only	0.681±0.017	0.545±0.015	0.653±0.019
Clinical-only	0.804±0.010	0.700±0.040	0.708±0.012
Early Fusion	0.813±0.011	**0.723±0.010**	**0.720±0.020**
Joint Fusion	0.809±0.009	0.710±0.029	0.715±0.020
Late Fusion	**0.836±0.017**	0.661±0.027	0.665±0.027
Multitask	0.689±0.037	0.558±0.065	0.627±0.047
Multitask Pretrained	0.701±0.024	0.581±0.029	0.637±0.029
Meta-learning	0.713±0.012	0.597±0.030	0.658±0.018

Observing the performance metrics in Table 5.2, we first notice that all methods and metrics generally have much higher variance as compared to results for detection of Gad lesion presence task in Table 5.1. While these tasks are different, high variance is often a sign of worse model generalization, potentially due to a drastically decreased dataset size and the underlying difficulty of predicting future disease activity. Comparing the uni-

modal methods, MRI-only approach performed significantly worse compared to clinical-only method across all metrics. We theorize that is due to the fact that baseline MRI scan is much less informative of future lesion and disease activity compared to the collection of clinical data. Logically, this makes sense because while MRI (single post-contrast FLAIR volume) can provide evidence of lesion activity at baseline, it is only a single data point that does not cover the overall clinical context of the patient. In contrast, clinical data (EDSS score, T2 lesion volume, Gad lesion count, etc.) provides multiple data points that are relevant to disease progression, and thus, are more informative when predicting future lesion activity.

We also notice a large discrepancy between methods where clinical information is provided "implicitly" (in transfer learning methods as meta-tasks or pretraining objectives) and "explicitly" (as model inputs in early, joint, and late fusion), with the latter showing much better performance. Similarly to the unimodal methods results, this points to the fact that the provided clinical data are more informative for successful prediction of future NE-T2 activity compared to the information contained in the MRI. We refer to this as *modality bias*. Methods where clinical information is part of the input to the model (early, joint, and late fusion) all performed similarly and outperformed all methods where clinical information was not provided as model input. Late fusion approach obtained highest ROC AUC, meanwhile early fusion method achieved highest PR AUC and F1 score. In addition, early and joint fusion marginally outperformed the unimodal methods across all metrics, again showing the benefits of utilizing multiple modalities. It is once again important to note that it is unknown whether the results are statistically significant, however, we believe that they can still demonstrate informative trends and are useful for intended analysis.

In order to confirm that the modality bias exists, we trained two identical early fusion methods while varying the imaging input. One method used real MRI data for training, meanwhile the other replaced the MRI input with random Gaussian noise with $\mu = 0$ and $\sigma^2 = 1$, but kept the same clinical data as input. ROC AUC performance on the validation

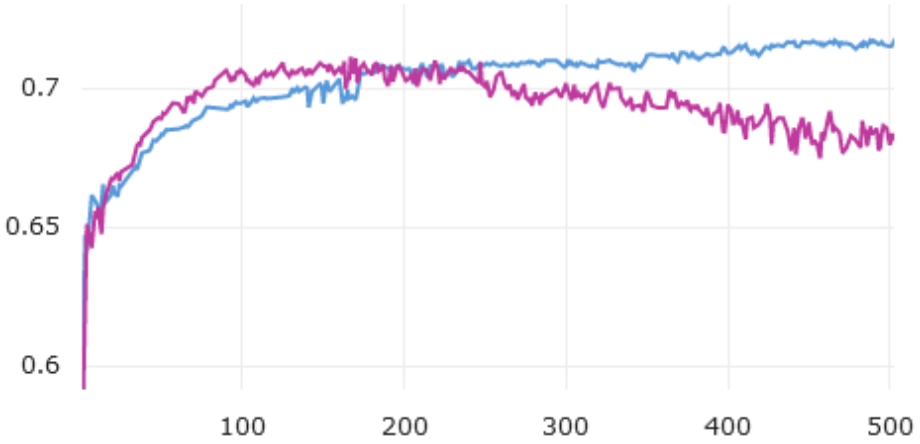

Figure 5.1: PR AUC performance on the validation set during model training using *early fusion* method for prediction of future NE-T2 lesion activity. Imaging input to the model using real MRI FLAIR sequence (purple) and random Gaussian noise (blue). Showing results for the first 500 epochs of training.

set with real MRI data can be observed in Figure 5.1, where using random Gaussian noise (red) reaches nearly identical performance as is using the real MRI scans (purple) for training. The fact that using random noise can achieve similar performance as the MRI demonstrates the existence of modality bias and the lack of useful information provided by the baseline FLAIR MRI sequence to the prediction of future NE-T2 activity.

As shown in Table 5.2, all of the transfer learning methods (multitask, multitask pretrained, and meta-learning) generally performed worse across all metrics than any methods using clinical data as part of the model input. This is expected due to the confirmed existing modality bias since these methods aim to improve performance by learning more informative imaging features and rely heavily on the information in the MRI. Nonetheless, both multitask methods achieved marginally better ROC AUC and PR AUC than MRI-only approach demonstrating the benefits of pretraining with clinical data. Interestingly, the proposed meta-learning approach performed better than both multitask as well as the MRI-only methods across all metrics. Given that MRI-only and meta-learning

methods have nearly identical model architectures and input data, the improved results indicate the benefits of meta-pretraining compared to training directly on the target task.

Estimation of ADAS-13 and MMSE cognitive scores

This section presents results of prediction of AD clinical score in Table 5.3 as RMSE mean and standard deviation (lower is better). As the first experiment, the models used T1-w MRI as the imaging input and non-imaging clinical data as the supporting clinical data (as mentioned in Chapter 4.4.2, specifically age, years of education, sex, and presence of APOE4 gene). In the later section, we also investigate the use of only image-derived as supporting clinical data.

Table 5.3: RMSE of various unimodal, multimodal, and meta-learning techniques for regression of ADAS-13 and MMSE clinical scores (using T1-w MRI and non-imaging clinical data). Performance is measured using 4-fold cross validation showing mean and standard deviation respectively. Gray are unimodal methods, cyan are multimodal methods, and green are transfer learning methods. Best results in bold, arrow direction indicates that lower value is better.

Method	**MMSE** $\downarrow$	**ADAS-13** $\downarrow$
MRI-only	1.843±0.119	7.323±0.179
Clinical-only	1.818±0.099	6.996±0.122
Early Fusion	**1.623±0.154**	**6.534±0.135**
Joint Fusion	1.652±0.132	6.608±0.145
Late Fusion	1.703±0.145	6.789±0.138
Multitask	1.787±0.185	7.101±0.143
Multitask Pretrained	1.813±0.176	7.204±0.151
Meta-learning	1.729±0.116	6.715±0.139

Observing the results for unimodal methods, clinical-only method achieved lower error than MRI-only method, more so for ADAS-13 score than MMSE. There is no large difference in mean RMSE between the two methods (compared to results in Table 5.2 for NE-T2 activity prediction), leading to the assumption that there is a low chance of modal-

ity bias and both modalities are similarly informative for the regression tasks. However, it is unknown if this is statistically significant.[1]

Multimodal methods (early, joint, and late fusion) once again demonstrated the benefits of utilizing multimodal data by generally achieving the lowest losses for both tasks, with early fusion performing the best and late fusion performing the worst out of the three. Transfer learning methods (multitask, multitask pretrained, and meta-learning) performed worse than the multimodal methods, however, all achieved lower loss than MRI-only method, indicating the minor benefits of CNN pretraining with the clinical data. Meta-learning method performed the best out of the three transfer learning methods, while also achieving lower loss for both tasks compared to the clinical-only method.

Similarly to the detection of Gad lesion presence task (Table 5.1), our proposed meta-learning approach achieved lower mean error than unimodal and transfer learning methods (multitask and multitask pretrained), yet it fell short of the performance of all multimodal methods. Given that in both tasks (lesion presence detection and AD score regression) there isn't a large performance difference between MRI-only and clinical-only methods (leading to assumption that there a low chance of modality bias), it was hypothesized that meta-learning would also perform on-par or better than multimodal methods, however, the results showed the opposite. Comparing the clinical data used between the two tasks, the ones used for Gad lesion presence detection are a mix of non-imaging (e.g. age, MS phenotype) and image-derived (e.g. T2 lesion volume) clinical features. In contrast, the clinical data used for AD score regression are all non-imaging markers. Believing that the absence of image-derived clinical data is related to the performance difference between multimodal and meta-learning methods, we investigated how selection of non-imaging or image-derived clinical data affects performance of the meta-learning approach on the target tasks. This is explored in the next section.

[1]Statistical significance test were not carried out due to computational and time complexity of the experiments, with a focus instead on very extensive experimentation and hyperparameter tuning, as well as the use of cross validation.

5.2 Selection and Effects of Supporting Clinical Data on Target Task Performance

In this section, we investigate the selection of supporting clinical data and its impact on the target task performance of our proposed meta-learning approach. Specifically, we compare performance of meta-learning approach that uses purely image-derived features (set of values that summarize important image-derived information) against using purely non-imaging features (as in Section 5.1.1) on the task of ADAS-13 and MMSE score regression. For non-imaging data, we used age, years of education, sex, and presence of APOE4 gene as was done in Section 5.1.1. First described in Chapter 4.4.2, we selected volumes of hippocampus, ventricles, entorhinal cortex, and whole brain as the image-derived features. Additionally, baseline methods were also trained using the two sets of supporting clinical data for comparison. All experiments are trained using an identical 4-fold cross validation setup, with the results presented in Table 5.4.

5.2.1 Results and Discussion

As presented in Table 5.4, clinical-only method achieved slightly lower loss on MMSE task (1.818 vs. 1.974 error) but relatively the same on ADAS-13 task when comparing non-imaging and image-derived clinical data. Similarly, multimodal methods performed largely the same when using either type of supporting clinical data, with the exception of early fusion which achieved marginally higher error using the image-derived markers. Interestingly, all transfer learning methods (multitask, multitask pretrained, and meta-learning) performed better with image-derived clinical data compared to non-imaging related. While it is unknown whether the findings are statistically significant, the proposed meta-learning approach also saw the largest improvement in terms of absolute error for both tasks among the transfer learning methods, and managed to achieve nearly the lowest error among all the methods. These findings lead us to believe that the meta-learning approach performs better when the meta-tasks used for pretraining are closely related to

Table 5.4: Comparing non-imaging (demographic; *Dem.*) and image-derived (*Img.*) supporting clinical data for ADAS-13 and MMSE regression in AD patients (identical T1-w MRI used for all experiments). Performance is measured using RMSE on 4-fold cross validation, showing mean and standard deviation respectively. Best results in bold, arrow direction indicates that lower value is better.

Method	**Clinical Data**		**MMSE ↓**	**ADAS-13 ↓**
	Dem.	**Img.**		
MRI-only			1.843±0.119	7.323±0.179
Clinical-only	✓		1.818±0.099	6.996±0.122
		✓	1.974±0.103	6.931±0.131
Early Fusion	✓		1.623±0.154	6.534±0.135
		✓	1.676±0.149	6.773±0.141
Joint Fusion	✓		1.652±0.132	6.608±0.145
		✓	1.650±0.129	6.598±0.127
Late Fusion	✓		1.703±0.145	6.789±0.138
		✓	1.694±0.106	6.698±0.104
Multitask	✓		1.787±0.185	7.101±0.143
		✓	1.709±0.164	6.753±0.121
Multitask Pretrained	✓		1.813±0.176	7.204±0.151
		✓	1.804±0.176	7.103±0.176
Meta-learning	✓		1.729±0.116	6.715±0.139
		✓	**1.621±0.102**	**6.219±0.130**

the information present in the imaging data (e.g. measurement of hippocampus volume that is directly visible in the MRI). While it would be logical to experiment with using all non-imaging and image-derived clinical features with the meta-learning approach, we do not perform such experiment due to significant additional computational resources required and complexity of hyperparameter tuning (elaborated on further in Section 5.4). Instead, we take a closer look at detection of Gad lesion presence experiments which use a combination of image-derived and non-imaging clinical features.

To further investigate the effects of non-imaging and image-derived clinical data on transfer learning performance, we explore whether the similar pattern can be found with the detection of Gad lesion presence task. Using the same set of supporting clinical data as in the original experiments in Section 5.1.1 (age, EDSS score, T2 lesion volume, and MS phenotype), we pretrained the CNN model to predict/regress each individual clini-

cal feature prior to finetuning on the target tasks. In this setup, the CNN learns imaging features that are only relevant to predicting a single clinical feature, as compared to predicting all clinical features in the case of multitask and meta-learning methods. Models are trained using the same hyperparameters and dataset as in Section 5.1.1 using 4-fold cross validation. ROC AUC and PR AUC metrics on the hold-out test set can be observed in Table 4.2.

Table 5.5: Test set ROC AUC AND PR AUC metrics on detection of Gad lesion presence after pretraining on varied individual clinical features. No pretraining (*MRI only*) achieved best performance, followed by pretraining on T2 lesion volume, MS phenotype, EDSS, and age. Performance is measured using 4-fold cross validation showing mean and standard deviation respectively. Results for meta-learning method using all of the supporting clinical features are also included for comparison. Best results in bold, arrow direction indicates that higher value is better.

Method	**ROC AUC** ↑	**PR AUC** ↑
MRI-only	0.822±0.006	0.763±0.008
T2 Volume	0.773±0.007	0.646±0.009
EDSS	0.746±0.007	0.606±0.008
Age	0.735±0.009	0.573±0.010
MS Phenotype	0.755±0.006	0.621±0.006
Meta-learning	**0.848±0.007**	**0.797±0.005**

Results in Table 5.5 show a clear separation in performance between different clinical features that were used for pretraining. As a baseline, model trained using only imaging information (*MRI-only*) managed to achieve the best performance compared to any of the single-feature pretraining methods, indicating that pretraining only on a single task is actually detrimental to the target task performance and can potentially lead to negative transfer. Between the pretraining methods, T2 lesion volume achieved better performance in both metrics over the other clinical markers, followed by MS phenotype classification task and EDSS score regression. Pretraining for age regression performed the worst and saw nearly no improvement over the course of finetuning. However, similarly to the AD experiments in Table 5.4, pretraining using an image-derived marker (T2 lesion volume)

still showed the best performance compared to non-imaging markers. Moreover, the next best performing clinical feature to pretrain on was MS phenotype. Within the context of detection of Gad lesion presence, knowing the MS disease phenotype (e.g. SPMS) is valuable since it is known that patients with RRMS are much more prone to developing Gad-enhanced lesions, which can also be visible in the MRI. As such, pretraining on classification of MS phenotype is indirectly related to the target task of detection presence of Gad lesions, as well as the information present in the MRI. At the same time, EDSS score [69] is a more subjective assessment of MS that measures overall patient ability and is not directly related to the target task. This is reflected by worse performance compared to pretraining on MS phenotype, however, still better than age regression.

5.3 Effects of Training Dataset Size on Target Task Performance

As mentioned in Chapter 2.3.3, meta-learning methods are often used in medical imaging for few-shot learning in order to combat lack of sufficient data. Typically, such methods [29, 127, 160] define meta-tasks as different datasets which are used to learn a single task (e.g. segmentation). In contrast, our proposed approach instead uses a single dataset but defines meta-tasks in terms of different learned tasks (e.g. regression, classification) from the same data. As a short exploratory study, we performed a set of experiments to understand how well our proposed meta-learning method performs across varying dataset sizes. As an example target task, we once again used detection of Gad lesion presence as in Section 5.1.1 due to the size of the dataset (3560 unique samples) and relatively good performance of the meta-learning approach in previous experiments. The MS dataset was split into 70% training, 10% validation, and 20% testing data splits. In order to simulate varying data regimes, the training split was then further reduced into 20%, 40%, 60%, 80%, and 100% splits of the overall 70% training splits. This was done by removing samples such that the distribution of samples with/out Gad lesion presence re-

mained identical (within 2-3%) across all splits in order to avoid a class distribution shift. Same methods as originally listed in Table 5.1 (for exception of clinical-only method) were trained using the dataset permutations above. Performance was evaluated using ROC AUC, PR AUC, and F1 score metrics on the hold-out test set.

5.3.1 Results and Discussion

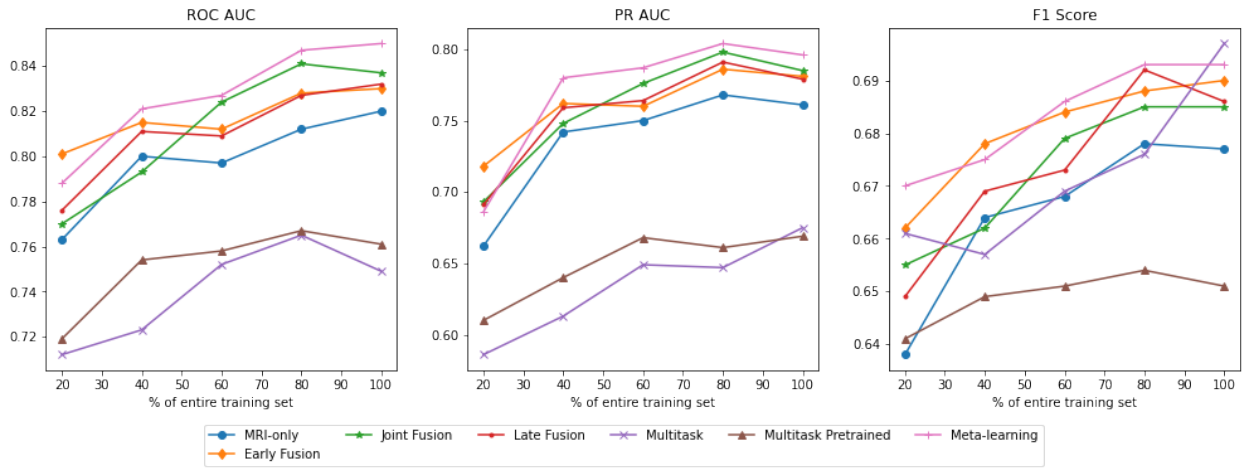

Figure 5.2: Effects of training dataset size across various image-based and multimodal methods on ROC AUC, PR AUC, and F1 score on the hold-out test set. Target task is detection of Gad lesion presence. All methods achieved higher performance as more training data is available; meta-learning approach generally achieved better performance across data regimes compared to other methods, with some exceptions.

Observing results in Figure 5.2, all methods generally tend to improve performance as more training data becomes available. This is expected behaviour given that more unique training samples allow the deep learning model to better learn the underlying data-label relationships as well as generalize on the unseen data [72]. Multimodal methods (early, joint, and late fusion) performed better in all metrics than MRI-only approach in nearly all data regimes. Surprisingly, among the transfer learning methods (multitask, multitask pretrained, and meta-learning) multitask and multitask pretrained methods performed significantly worse compared to other methods across all the metrics, for the exception of

multimodal method's F1 score. We believe that this is due to the fact that using MRI to predict several clinical features at once (multitask), where each permutation of the clinical features can be treated as a unique ground truth label, is suffering from an even worse lack of data compared to when the same input MRI is used for prediction of a single label (e.g. Gad lesion presence detection). In contrast, the proposed meta-learning approach outperformed (in some cases marginally) all other methods across all metrics and training dataset sizes for the exception of 20% training data split case. Interestingly, meta-learning generally achieved the same ROC AUC, PR AUC, and in some cases F1 score as the next best method with 20% less data available (with some exceptions). We believe that the performance of meta-learning in lower data regimes can be attributed to meta-pretraining allowing for a larger number of unique gradient updates compared to training a model directly on the target task. For example, in a 20% training dataset case, there are 498 unique MRI samples with the respective target label (ignoring additional samples due to augmentations). When pretraining with meta-learning, the same 498 MRI samples also gain four additional and unique target labels (the meta-tasks) which provide additional training guidance for the model. In essence, the input data gap problem is mitigated through additional training supervision with extra ground truth labels (the meta-tasks). However, this once again relies on the assumption that the learned shared knowledge from the meta-tasks is related and beneficial to the target task. While this section presents a potential direction for future research, it is only a shallow exploration with experiments done on a single permutation of the dataset (e.g. no cross-validation, unknown statistical significance), and would require further experimentation to confirm the initial findings.

5.4 Limitations

Throughout experimentation, we identified a number of limitations of our proposed meta-learning approach. Compared to single-stage training methods such as multi-task, multi-modal, and unimodal methods, the meta-learning approach has separate pretraining and

finetuning stages. This introduces a number of challenges with the selection of supporting clinical features and hyperparameter tuning. While there are guidelines for selecting which clinical features to use as meta-tasks, for example, the need to be related to the information in the MRI and the target task, it is not guaranteed that the learning a shared representation with meta-pretraining will result in better performance on the target task after finetuning. An example case of this was demonstrated with modality bias in experiments in Section 5.1.1. This is due to the fact that there is no feedback signal such as a loss gradient from performance on the target task to the meta-pretraining process. Additionally, this severely limits which clinical features can be used even if they are potentially beneficial to the learning of the target task, but not in a shared learning environment with other meta-tasks. Lastly, our approach is sensitive to cases with severe modality bias (see experiments in Section 5.1.1), where MRI was largely uninformative of the target task compared to clinical data. At the same time, multimodal methods do not have these limitations as they can adjust the contribution of each of the input sources directly when learning the target task, with the model automatically learning the optimal feature selection by itself.

Another downside of the proposed method is that the two-stage training approach makes hyperparameter tuning and validation much more complicated. Essentially, the hyperparameter tuning process is separated into two parts: meta-pretraining tuning and target task adaptation tuning, with the goal of maximizing validation performance on the target task. In addition, there are extra hyperparameters specific to the meta-learning process that need to be taken into account. For example, our implementation assumes that all meta-tasks are equally important during meta-pretraining stage, however, that might not always be the case. As such, selecting appropriate weights (ω_n in Algorithm 1) for the meta-tasks becomes an additional consideration to take into account. In summary, meta-learning approach requires a lot more user effort and computational time compared to single-stage methods.

Our proposed approach is also computationally and resource expensive. During meta-pretraining using batch version of Reptile, an individual meta-model is trained on each clinical feature leading to N parallel models when using a set of N clinical features. Depending on the implementation, the meta-pretraining stage either requires N-fold more computation time (each meta-model trained in series per training step), or N-fold more compute resources (if trained in parallel, e.g. one GPU per model) when compared to other multimodal fusion methods. This makes using large number of clinical features (and by extension, meta-tasks) computationally unreasonable. One way to solve this issue would be to sequentially randomly sample individual meta-tasks during pretraining, however, this increases the training time and raises question of how to ensure fair meta-task representation, which is an on-going research area in itself [14, 150].

Lastly, it is important to note the effects of dataset size on our experiments. The subset of MS dataset that was used contained over 3500 patient samples, and is considered as a large dataset in the medical imaging domain. For comparison, other large brain MRI datasets such as BrATS 2020 [91] and a variation of a broader ADNI dataset [87] contained 2640 and 1886 patient samples respectively. However, these datasets are considered very small in the context of a DL system which are known to perform better with large amounts of data (e.g. a small computer vision dataset for benchmarking DL methods, Tiny ImageNet [71], contains 100,000 images). In addition to 3D nature and structure complexity within the MRI volumes, this lack of data can hamper training and model performance of an ML system.

5.5 Summary

This chapter presented experimental results and comparison between the proposed meta-learning and existing unimodal, multimodal, and transfer learning methods trained for three medical imaging tasks. First, performance metrics of multimodal methods compared to unimodal methods generally showed benefits of utilizing both the MRI and

clinical data for all of the three target tasks, with minor exceptions. For two of them (detection of Gad lesion presence and estimation of ADAS-13 and MMSE cognitive scores) meta-learning performed the best across all methods. For the task of future NE-T2 lesion activity prediction, experimental results identified a significant modality bias towards the clinical data, which is believed to be the cause of poor performance by the transfer learning methods. This chapter also presented an ablation study comparing use of image-derived and non-imaging related clinical features with the proposed meta-learning approach. Experimental results demonstrated that meta-learning methods benefit the most when the supporting clinical markers are closely related to the target task as well as the content found in the MRI sequence itself. Lastly, a short exploratory study illustrated the effectiveness of meta-learning method in dealing with lack of training data, where the results showed meta-learning outperforming other methods in all but the lowest data regimes across nearly all metrics. While the statistical significance of the above results is unknown, we believe the results are sufficient to identify the discussed trends. Lastly, this chapter discussed some of the downfalls of our proposed approach, in particular, complexity of hyperparameter tuning, selection of supporting clinical data, and high computational requirements.

Chapter 6

Conclusion

This thesis presented a meta-learning approach to improving deep learning model performance when clinical information is provided in addition to medical imaging data. First, we illustrated the importance of using multimodal data for correct interpretation of medical images due to the problem of visual ambiguities in the imaging data, and hypothesized that addition of clinical information provides valuable clinical context and patient history that can improve model performance compared to imaging-only methods. Various multimodal and transfer learning approaches that make use of multiple data modalities were then introduced, and we listed a number of their successful existing applications within the medical imaging domain. We then presented our adaptation of Reptile meta-learning algorithm for pretraining image-based CNN models with tabular clinical features, followed by finetuning on the desired target task. To evaluate our method, we trained all methods to perform three example medical imaging tasks: 1) detection of Gad lesion presence, 2) prediction of future NE-T2 lesion activity, and 3) regression of ADAS-13 and MMSE cognitive scores. Experimental results in Chapter 5.1 demonstrated that our proposed meta-learning approach performed not only better than imaging-only methods for tasks #1 and #3 in select metrics, but also marginally outperforming multimodal and other transfer learning methods in some cases. Interestingly, we also identified a case of modality bias in the prediction of future NE-T2 lesion activ-

ity task, where selected clinical feature set was much more informative compared to the information in the MRI data. However, even in this case, the proposed meta-learning approach achieved better metrics than the imaging-only method, once again showing the benefits of meta-pretraining with clinical context. While it is unknown whether these findings are statistically significant, they do show trends that are convincing enough to warrant further research into applications of meta-learning in the medical imaging domain.

Experimental results of supporting clinical data ablation (5.2) demonstrated that clinical features that are used as meta-tasks during meta-pretraining should be the ones that are closely related to 1) the target task and 2) the information in the MRI itself in order to achieve the best target task performance with our proposed method. This was inline with our original assumptions and was specifically observed with the tasks of Gad lesion presence detection in MS and cognitive score regression in AD. Lastly, results of the short dataset size ablation experiments in Chapter 5.3 showed impressive robustness of our proposed approach by generally achieving the best metric performance compared to unimodal, multimodal, and other transfer learning methods across nearly all data regimes. In summary, we presented experimental evidence of the merits of the proposed meta-learning approach for making use of clinical information with medical image-based DL models, a guideline on selection of the supporting clinical feature set, and an exploratory look at robustness across different data regimes.

One promising future research direction is to investigate how to improve the meta-pretraining procedure with regards to performance on the target task. For example, determining the loss on the target task during meta-pretraining stage and using it to adjust the meta-model parameter updates, or to choose the duration of the pretraining process (e.g. stopping meta-pretraining when target loss is at the lowest). Another possible future extension of the work is to explore our proposed use of meta-learning for improving parameter initialization of DL models used for medical image segmentation (e.g. brain lesion or white matter). For example, a segmentation model (e.g. UNet [115]) can be

pretrained to predict relevant clinical information (e.g. disease stage) by using the meta-learning scheme. Following that, the model can be finetuned for segmentation, and in theory could potentially achieve more accurate segmentation given that the model has prior knowledge of the clinical context (e.g. later disease stage relates to more lesions being present).

Bibliography

[1] ABDULNABI, A. H., WANG, G., LU, J., AND JIA, K. Multi-task cnn model for attribute prediction. *Trans. Multi. 17*, 11 (nov 2015), 1949–1959.

[2] AHMED, I., QUINONES-GRUEIRO, M., AND BISWAS, G. Complementary meta-reinforcement learning for fault-adaptive control. *Annual Conference of the PHM Society 12* (11 2020), 8.

[3] ALBAWI, S., MOHAMMED, T. A., AND AL-ZAWI, S. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (2017), pp. 1–6.

[4] BALDI, P. Gradient descent learning algorithm overview: a general dynamical systems perspective. *IEEE Transactions on Neural Networks 6*, 1 (1995), 182–195.

[5] BARZEGAR, M., SHAYGANNEJAD, V., MIRMOSAYYEB, O., AND AFSHARI, A. Progression to secondary progressive multiple sclerosis and its early risk factors: A population-based study (2171). *Neurology 94*, 15 Supplement (2020).

[6] BAUMGART, M., SNYDER, H. M., CARRILLO, M. C., FAZIO, S., KIM, H., AND JOHNS, H. Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimer's &amp Dementia 11*, 6 (May 2015), 718–726.

[7] BHAGWAT, N., VIVIANO, J. D., VOINESKOS, A. N., AND AND, M. M. C. Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLOS Computational Biology 14*, 9 (Sept. 2018), e1006376.

[8] BISSOTO, A., VALLE, E., AND AVILA, S. Debiasing skin lesion datasets and models? not so fast. pp. 3192–3201.

[9] BLANCHARD, M., AND BENNOUNA, M. A. The representation power of neural networks: Breaking the curse of dimensionality. *ArXiv abs/2012.05451* (2020).

[10] BOONN, W. W., AND LANGLOTZ, C. P. Radiologist use of and perceived need for patient data access. *Journal of Digital Imaging 22*, 4 (May 2008), 357–362.

[11] CALHOUN, V. D., AND SUI, J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 1*, 3 (May 2016), 230–244.

[12] CARLSSON, C. M., GLEASON, C. E., PUGLIELLI, L., AND ASTHANA, S. *Dementia Including Alzheimer Disease*. McGraw-Hill Education, New York, NY, 2017.

[13] CHEN, C., HAN, D., AND WANG, J. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access 8* (2020), 35662–35671.

[14] CHEN, Y., ZHANG, S., AND KIAN HSIANG LOW, B. Near-optimal task selection for meta-learning with mutual information and online variational bayesian unlearning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* (28–30 Mar 2022), G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., vol. 151 of *Proceedings of Machine Learning Research*, PMLR, pp. 9091–9113.

[15] CHUPIN, M., GÉRARDIN, E., CUINGNET, R., BOUTET, C., LEMIEUX, L., LEHÉRICY, S., BENALI, H., GARNERO, L., AND AND, O. C. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus 19*, 6 (June 2009), 579–587.

[16] COLLINS, D. L., NEELIN, P., PETERS, T., AND EVANS, A. C. Automatic 3d inter-subject registration of mr volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography 18* (1994), 192–205.

[17] CONSORTIUM, T. M. Project monai, 2020.

[18] CUI, H., RADOSAVLJEVIC, V., CHOU, F.-C., LIN, T.-H., NGUYEN, T., HUANG, T.-K., SCHNEIDER, J., AND DJURIC, N. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)* (2019), pp. 2090–2096.

[19] DADAR, M., FONOV, V. S., AND COLLINS, D. L. A comparison of publicly available linear MRI stereotaxic registration techniques. *NeuroImage 174* (July 2018), 191–200.

[20] DING, B., QIAN, H., AND ZHOU, J. Activation functions and their characteristics in deep neural networks. In *2018 Chinese Control And Decision Conference (CCDC)* (2018), pp. 1836–1841.

[21] DOSHI, A.M., H. C. G. L. E. A. Impact of patient questionnaires on completeness of clinical information and identification of causes of pain during outpatient abdominopelvic ct interpretation. *Abdom Radiol 42* (2017), 2946—-2950.

[22] DOYLE, A., ELLIOTT, C., KARIMAGHALOO, Z., SUBBANNA, N., ARNOLD, D. L., AND ARBEL, T. Lesion detection, segmentation and prediction in multiple sclerosis clinical trials. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2018, pp. 15–28.

[23] DOYLE, A., PRECUP, D., ARNOLD, D. L., AND ARBEL, T. Predicting future disease activity and treatment responders for multiple sclerosis patients using a bag-of-lesions brain representation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing, 2017, pp. 186–194.

[24] DUC, N. T., RYU, S., QURESHI, M. N. I., CHOI, M., LEE, K. H., AND LEE, B. 3d-deep learning based automatic diagnosis of alzheimer's disease with joint MMSE prediction using resting-state fMRI. *Neuroinformatics 18*, 1 (May 2019), 71–86.

[25] DURSO-FINLEY, J., FALET, J.-P., NICHYPORUK, B., ARNOLD, D., AND ARBEL, T. Personalized prediction of future lesion activity and treatment effect in multiple sclerosis from baseline mri.

[26] ERHAN, D., BENGIO, Y., COURVILLE, A., MANZAGOL, P.-A., VINCENT, P., AND BENGIO, S. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research 11*, 19 (2010), 625–660.

[27] FALLAH, A., MOKHTARI, A., AND OZDAGLAR, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (26–28 Aug 2020), S. Chiappa and R. Calandra, Eds., vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 1082–1092.

[28] FALLAH, A., MOKHTARI, A., AND OZDAGLAR, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2020), NIPS'20, Curran Associates Inc.

[29] FARSHAD, A., MAKAREVICH, A., BELAGIANNIS, V., AND NAVAB, N. MetaMedSeg: Volumetric meta-learning for few-shot organ segmentation. In *Domain Adaptation and Representation Transfer*. Springer Nature Switzerland, 2022, pp. 45–55.

[30] FEDERATION, M. I. *Atlas of MS 3rd edition Part 1: Mapping Multiple Sclerosis Around the World*, 3 ed. MS International Federation, 2020.

[31] FINN, C., ABBEEL, P., AND LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine*

Learning (06–11 Aug 2017), D. Precup and Y. W. Teh, Eds., vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 1126–1135.

[32] FOLSTEIN, M. F., FOLSTEIN, S. E., AND MCHUGH, P. R. "mini-mental state". *Journal of Psychiatric Research 12*, 3 (Nov. 1975), 189–198.

[33] FREEDMAN, M. S., DEVONSHIRE, V., DUQUETTE, P., GIACOMINI, P. S., GIULIANI, F., LEVIN, M. C., MONTALBAN, X., MORROW, S. A., OH, J., ROTSTEIN, D., AND YEH, E. A. Treatment optimization in multiple sclerosis: Canadian MS working group recommendations. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques 47*, 4 (Apr. 2020), 437–455.

[34] FRISONI, G. B., FOX, N. C., JACK, C. R., SCHELTENS, P., AND THOMPSON, P. M. The clinical use of structural MRI in alzheimer disease. *Nature Reviews Neurology 6*, 2 (Feb. 2010), 67–77.

[35] GAJ, S., ONTANEDA, D., AND NAKAMURA, K. Automatic segmentation of gadolinium-enhancing lesions in multiple sclerosis using deep learning from clinical MRI. *PLOS ONE 16*, 9 (Sept. 2021), e0255939.

[36] GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M., AND LEMPITSKY, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res. 17*, 1 (jan 2016), 2096–2030.

[37] GEIRHOS, R., JACOBSEN, J.-H., MICHAELIS, C., ZEMEL, R., BRENDEL, W., BETHGE, M., AND WICHMANN, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence 2*, 11 (Nov. 2020), 665–673.

[38] GENOVESE, A. V., HAGEMEIER, J., BERGSLAND, N., JAKIMOVSKI, D., DWYER, M. G., RAMASAMY, D. P., LIZARRAGA, A. A., HOJNACKI, D., KOLB, C., WEINSTOCK-GUTTMAN, B., AND ZIVADINOV, R. Atrophied brain t2 lesion volume at MRI is associated with disability progression and conversion to secondary progressive multiple sclerosis. *Radiology 293*, 2 (Nov. 2019), 424–433.

[39] GESSERT, N., BENGS, M., KRÜGER, J., OPFER, R., OSTWALDT, A.-C., MANOGARAN, P., SCHIPPLING, S., AND SCHLAEFER, A. 4d deep learning for multiple sclerosis lesion activity segmentation, 2020.

[40] GOGATE, M., ADEEL, A., AND HUSSAIN, A. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (2017), pp. 1–6.

[41] GOLDENBERG, M. M. Multiple sclerosis review. *P & T : a peer-reviewed journal for formulary management 37 3* (2012), 175–84.

[42] GORDON, J., BRONSKILL, J., BAUER, M., NOWOZIN, S., AND TURNER, R. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations* (2019).

[43] HARDING, K. E., LIANG, K., COSSBURN, M. D., INGRAM, G., HIRST, C. L., PICKERSGILL, T. P., TE WATER NAUDE, J., WARDLE, M., BEN-SHLOMO, Y., AND ROBERTSON, N. P. Long-term outcome of paediatric-onset multiple sclerosis: a population-based study. *Journal of Neurology, Neurosurgery & Psychiatry 84*, 2 (2013), 141–147.

[44] HE, K., ZHANG, X., REN, S., AND SUN, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1026–1034.

[45] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[46] HEATON, J. An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016* (2016), pp. 1–6.

[47] HECHT-NIELSEN. Theory of the backpropagation neural network. In *International 1989 Joint Conference on Neural Networks* (1989), pp. 593–605 vol.1.

[48] HECKER, S., DAI, D., AND GOOL, L. V. End-to-end learning of driving models with surround-view cameras and route planners. In *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 449–468.

[49] HERMESSI, H., MOURALI, O., AND ZAGROUBA, E. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing 183* (2021), 108036.

[50] HOSPEDALES, T., ANTONIOU, A., MICAELLI, P., AND STORKEY, A. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 9 (2022), 5149–5169.

[51] HUANG, S.-C., PAREEK, A., SEYYEDI, S., BANERJEE, I., AND LUNGREN, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine 3*, 1 (Oct. 2020).

[52] HUANG, S.-C., PAREEK, A., ZAMANIAN, R., BANERJEE, I., AND LUNGREN, M. P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports 10*, 1 (Dec. 2020).

[53] HUANG, Z.-A., HU, Y., LIU, R., XUE, X., ZHU, Z., SONG, L., AND TAN, K. C. Federated multi-task learning for joint diagnosis of multiple mental disorders on mri scans. *IEEE Transactions on Biomedical Engineering* (2022), 1–12.

[54] HYUN, S. H., AHN, M. S., KOH, Y. W., AND LEE, S. J. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clinical Nuclear Medicine 44*, 12 (Oct. 2019), 956–960.

[55] ILIEVSKI, I., AND FENG, J. Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems 30* (2017).

[56] JACK, C. R., KNOPMAN, D. S., JAGUST, W. J., PETERSEN, R. C., WEINER, M. W., AISEN, P. S., SHAW, L. M., VEMURI, P., WISTE, H. J., WEIGAND, S. D., LESNICK, T. G., PANKRATZ, V. S., DONOHUE, M. C., AND TROJANOWSKI, J. Q. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology 12*, 2 (Feb. 2013), 207–216.

[57] JAISWAL, A., BABU, A. R., ZADEH, M. Z., BANERJEE, D., AND MAKEDON, F. A survey on contrastive self-supervised learning. *Technologies 9*, 1 (2021).

[58] JAYARATNE, M., ALAHAKOON, D., SILVA, D. D., AND YU, X. Bio-inspired multisensory fusion for autonomous robots. In *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society* (Oct. 2018), IEEE.

[59] JING, L., AND TIAN, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 11 (2021), 4037–4058.

[60] KARIMAGHALOO, Z., RIVAZ, H., ARNOLD, D. L., COLLINS, D. L., AND ARBEL, T. Adaptive voxel, texture and temporal conditional random fields for detection of gad-enhancing multiple sclerosis lesions in brain MRI. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Springer Berlin Heidelberg, 2013, pp. 543–550.

[61] KARIMAGHALOO, Z., SHAH, M., FRANCIS, S. J., ARNOLD, D. L., COLLINS, D. L., AND ARBEL, T. Detection of gad-enhancing lesions in multiple sclerosis using conditional random fields. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part III 13* (2010), Springer, pp. 41–48.

[62] KAUR, B., LEMAÎTRE, P., MEHTA, R., SEPAHVAND, N. M., PRECUP, D., ARNOLD, D., AND ARBEL, T. Improving pathological structure segmentation via transfer learning across diseases. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer International Publishing, 2019, pp. 90–98.

[63] KAWAHARA, J., DANESHVAR, S., ARGENZIANO, G., AND HAMARNEH, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics 23*, 2 (2019), 538–546.

[64] KHANDELWAL, P., AND YUSHKEVICH, P. Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer International Publishing, 2020, pp. 73–84.

[65] KIM, J., BASAK, J. M., AND HOLTZMAN, D. M. The role of apolipoprotein e in alzheimer's disease. *Neuron 63*, 3 (Aug. 2009), 287–303.

[66] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Y. Bengio and Y. LeCun, Eds.

[67] KOCH-HENRIKSEN, N., AND SØRENSEN, P. S. The changing demographic pattern of multiple sclerosis epidemiology. *The Lancet Neurology 9*, 5 (May 2010), 520–532.

[68] KRISHNAN, A. P., SONG, Z., CLAYTON, D., GAETANO, L., JIA, X., DE CRESPIGNY, A., BENGTSSON, T., AND CARANO, R. A. D. Joint MRI t1 unenhancing and contrast-enhancing multiple sclerosis lesion segmentation with deep learning in OPERA trials. *Radiology 302*, 3 (Mar. 2022), 662–673.

[69] KURTZKE, J. F. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology 33*, 11 (Nov. 1983), 1444–1444.

[70] LAHAT, D., ADALÝ, T., AND JUTTEN, C. Challenges in multimodal data fusion. In *2014 22nd European Signal Processing Conference (EUSIPCO)* (2014), pp. 101–105.

[71] LE, Y., AND YANG, X. Tiny imagenet visual recognition challenge.

[72] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *Nature 521*, 7553 (May 2015), 436–444.

[73] LEI, B., LIANG, E., YANG, M., YANG, P., ZHOU, F., TAN, E.-L., LEI, Y., LIU, C.-M., WANG, T., XIAO, X., AND WANG, S. Predicting clinical scores for alzheimer's disease based on joint and deep learning. *Expert Systems with Applications 187* (Jan. 2022), 115966.

[74] LI, H., AND FAN, Y. Early prediction of alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Apr. 2019), IEEE.

[75] LI, J., ZHU, G., HUA, C., FENG, M., BENNAMOUN, B., LI, P., LU, X., SONG, J., SHEN, P., XU, X., MEI, L., ZHANG, L., SHAH, S. A. A., AND BENNAMOUN. A systematic collection of medical image datasets for deep learning. *ArXiv abs/2106.12864* (2021).

[76] LI, M., ZHANG, Y., HAN, D., AND ZHOU, M. Meta-IP: An imbalanced processing model based on meta-learning for IT project extension forecasts. *Mathematical Problems in Engineering 2022* (Sept. 2022), 1–11.

[77] LI, T., SANJABI, M., BEIRAMI, A., AND SMITH, V. Fair resource allocation in federated learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (2020), OpenReview.net.

[78] LI, X., CHEN, H., QI, X., DOU, Q., FU, C.-W., AND HENG, P.-A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging 37*, 12 (Dec. 2018), 2663–2674.

[79] LI, X., YU, L., JIN, Y., FU, C.-W., XING, L., AND HENG, P.-A. Difficulty-aware meta-learning for rare disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing, 2020, pp. 357–366.

[80] LI, Y., WANG, J., YE, J., AND REDDY, C. K. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, p. 1715–1724.

[81] LIANG, J., JIANG, L., CAO, L., LI, L.-J., AND HAUPTMANN, A. G. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6135–6143.

[82] LIU, Q., DOU, Q., AND HENG, P.-A. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing, 2020, pp. 475–485.

[83] LIU, Z., MIAO, Z., ZHAN, X., WANG, J., GONG, B., AND YU, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[84] MANDAL, P. K., MAHAJAN, R., AND DINOV, I. D. Structural brain atlases: Design, rationale, and applications in normal and pathological cohorts. *Journal of Alzheimer's Disease 31*, s3 (Sept. 2012), S169–S188.

[85] MANJÓN, J. V., COUPÉ, P., MARTÍ-BONMATÍ, L., COLLINS, D. L., AND ROBLES, M. Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging 31*, 1 (Dec. 2009), 192–203.

[86] MAO, K., LU, D., E, D., AND TAN, Z. A case study on attribute recognition of heated metal mark image using deep convolutional neural networks. *Sensors 18* (06 2018), 1871.

[87] MARINESCU, R. V., OXTOBY, N. P., YOUNG, A. L., BRON, E. E., TOGA, A. W., WEINER, M. W., BARKHOF, F., FOX, N. C., GOLLAND, P., KLEIN, S., AND ALEXANDER, D. C. TADPOLE challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine*. Springer International Publishing, 2019, pp. 1–10.

[88] MATTA, S., LAMARD, M., CONZE, P.-H., GUILCHER, A. L., RICQUEBOURG, V., BENYOUSSEF, A.-A., MASSIN, P., ROTTIER, J.-B., COCHENER, B., AND QUELLEC, G. Meta learning for anomaly detection in fundus photographs. In *Meta Learning With Medical Imaging and Health Informatics Applications*. Elsevier, 2023, pp. 301–329.

[89] MCDONALD, R. J., SCHWARTZ, K. M., ECKEL, L. J., DIEHN, F. E., HUNT, C. H., BARTHOLMAI, B. J., ERICKSON, B. J., AND KALLMES, D. F. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology 22*, 9 (Sept. 2015), 1191–1198.

[90] MEHTA, R., CHRISTINCK, T., NAIR, T., BUSSY, A., PREMASIRI, S., COSTANTINO, M., CHAKRAVARTHY, M. M., ARNOLD, D. L., GAL, Y., AND ARBEL, T. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Transactions on Medical Imaging 41*, 2 (2022), 360–373.

[91] MEHTA, R., FILOS, A., BAID, U., SAKO, C., MCKINLEY, R., REBSAMEN, M., DATWYLER, K., MEIER, R., RADOJEWSKI, P., MURUGESAN, G. K., NALAWADE, S.,

GANESH, C., WAGNER, B., YU, F. F., FEI, B., MADHURANTHAKAM, A. J., MALDJIAN, J. A., DAZA, L., GOMEZ, C., ARBELAEZ, P., DAI, C., WANG, S., REYNAUD, H., MO, Y.-H., ANGELINI, E., GUO, Y., BAI, W., BANERJEE, S., PEI, L.-M., AK, M., ROSAS-GONZALEZ, S., ZEMMOURA, I., TAUBER, C., VU, M. H., NYHOLM, T., LOFSTEDT, T., BALLESTAR, L. M., VILAPLANA, V., MCHUGH, H., TALOU, G. M., WANG, A., PATEL, J., CHANG, K., HOEBEL, K., GIDWANI, M., ARUN, N., GUPTA, S., AGGARWAL, M., SINGH, P., GERSTNER, E. R., KALPATHY-CRAMER, J., BOUTRY, N., HUARD, A., VIDYARATNE, L., RAHMAN, M. M., IFTEKHARUDDIN, K. M., CHAZALON, J., PUYBAREAU, E., TOCHON, G., MA, J., CABEZAS, M., LLADO, X., OLIVER, A., VALENCIA, L., VALVERDE, S., AMIAN, M., SOLTANINEJAD, M., MYRONENKO, A., HATAMIZADEH, A., FENG, X., DOU, Q., TUSTISON, N., MEYER, C., SHAH, N. A., TALBAR, S., WEBER, M.-A., MAHAJAN, A., JAKAB, A., WIEST, R., FATHALLAH-SHAYKH, H. M., NAZERI, A., MILCHENKO1, M., MARCUS, D., KOTROTSOU, A., COLEN, R., FREYMANN, J., KIRBY, J., DAVATZIKOS, C., MENZE, B., BAKAS, S., GAL, Y., AND ARBEL, T. Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation - analysis of ranking scores and benchmarking results.

[92] MYERS-COLET, C., SCHROETER, J., ARNOLD, D. L., AND ARBEL, T. Heatmap regression for lesion detection using pointwise annotations. In *Medical Image Learning with Limited and Noisy Data*. Springer Nature Switzerland, 2022, pp. 3–12.

[93] NADEEM, M. W., GOH, H. G., ALI, A., HUSSAIN, M., KHAN, M. A., AND A/P PONNUSAMY, V. Bone age assessment empowered with deep learning: A survey, open research challenges and future directions. *Diagnostics 10*, 10 (Oct. 2020), 781.

[94] NANA, A., RUTH, A. M., CHRISTINA, B., ROCHELLE, G., DOUGLAS, G. M., RON, W., PHILIPPE, F., JULIE, B., KAREN, T., AND KIM, R. Multiple sclerosis in canada 2011 to 2031: results of a microsimulation modelling study of epidemiological and

economic impacts. *Health promotion and chronic disease prevention in Canada: research, policy and practice 37*, 2 (2017), 37.

[95] NGO, D.-K., TRAN, M.-T., KIM, S.-H., YANG, H.-J., AND LEE, G.-S. Multi-task learning for small brain tumor segmentation from mri. *Applied Sciences 10*, 21 (2020).

[96] NGUYEN, A., NGUYEN, N., TRAN, K., TJIPUTRA, E., AND TRAN, Q. D. Autonomous navigation in complex environments with deep multimodal fusion network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Oct. 2020), IEEE.

[97] NICHOL, A., ACHIAM, J., AND SCHULMAN, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).

[98] OBAMUYIDE, A., AND VLACHOS, A. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 5873–5879.

[99] OH, S.-I., AND KANG, H.-B. Object detection and classification by decision-level fusion for intelligent vehicle systems. *Sensors 17*, 12 (Jan. 2017), 207.

[100] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., DESMAISON, A., KOPF, A., YANG, E., DEVITO, Z., RAISON, M., TEJANI, A., CHILAMKURTHY, S., STEINER, B., FANG, L., BAI, J., AND CHINTALA, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[101] PATRO, B., PATEL, S., AND NAMBOODIRI, V. Robust explanations for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (March 2020).

[102] PEREZ-RUA, J.-M., ZHU, X., HOSPEDALES, T. M., AND XIANG, T. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13846–13855.

[103] PERSON, M., JENSEN, M., SMITH, A. O., AND GUTIERREZ, H. Multimodal fusion object detection system for autonomous vehicles. *Journal of Dynamic Systems, Measurement, and Control 141*, 7 (May 2019).

[104] PETERSEN, R. C., AISEN, P. S., BECKETT, L. A., DONOHUE, M. C., GAMST, A. C., HARVEY, D. J., JACK, C. R., JAGUST, W. J., SHAW, L. M., TOGA, A. W., TROJANOWSKI, J. Q., AND WEINER, M. W. Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology 74*, 3 (Dec. 2009), 201–209.

[105] PRABHU, V., KANNAN, A., RAVURI, M., CHAPLAIN, M., SONTAG, D., AND AMATRIAIN, X. Few-shot learning for dermatological disease diagnosis. In *Proceedings of the 4th Machine Learning for Healthcare Conference* (09–10 Aug 2019), F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 106 of *Proceedings of Machine Learning Research*, PMLR, pp. 532–552.

[106] QIU, S., CHANG, G. H., PANAGIA, M., GOPAL, D. M., AU, R., AND KOLACHALAMA, V. B. Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's &amp Dementia: Diagnosis, Assessment &amp Disease Monitoring 10*, 1 (Jan. 2018), 737–749.

[107] RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., KRUEGER, G., AND SUTSKEVER, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (2021), M. Meila and T. Zhang, Eds., vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 8748–8763.

[108] RADU, V., LANE, N. D., BHATTACHARYA, S., MASCOLO, C., MARINA, M. K., AND KAWSAR, F. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (New York, NY, USA, 2016), UbiComp '16, Association for Computing Machinery, p. 185–188.

[109] RAMACHANDRAM, D., LISICKI, M., SHIELDS, T., AMER, M., AND TAYLOR, G. Structure optimization for deep multimodal fusion networks using graph-induced kernels.

[110] RAMACHANDRAM, D., AND TAYLOR, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine 34*, 6 (2017), 96–108.

[111] REDA, I., KHALIL, A., ELMOGY, M., EL-FETOUH, A. A., SHALABY, A., EL-GHAR, M. A., ELMAGHRABY, A., GHAZAL, M., AND EL-BAZ, A. Deep learning role in early diagnosis of prostate cancer. *Technology in Cancer Research &amp Treatment 17* (Jan. 2018), 153303461877553.

[112] REED, R. Pruning algorithms-a survey. *IEEE Transactions on Neural Networks 4*, 5 (1993), 740–747.

[113] REN, M., TRIANTAFILLOU, E., RAVI, S., SNELL, J., SWERSKY, K., TENENBAUM, J. B., LAROCHELLE, H., AND ZEMEL, R. S. Meta-learning for semi-supervised few-shot classification. *ArXiv abs/1803.00676* (2018).

[114] RODRIGUES, R., MADEIRA, R. N., CORREIA, N., FERNANDES, C., AND RIBEIRO, S. Multimodal web based video annotator with real-time human pose estimation. In *International Conference on Intelligent Data Engineering and Automated Learning* (2019), Springer, pp. 23–30.

[115] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. vol. 9351, pp. 234–241.

[116] ROVIRA, A., TINTORÈ, M., ALVAREZ-CERMENO, J. C., IZQUIERDO, G., AND PRIETO, J. Recommendations for using and interpreting magnetic resonance imaging in multiple sclerosis [article in spanish]. *Neurología (Barcelona, Spain) 25* (05 2010), 248–65.

[117] RUDER, S. An overview of multi-task learning in deep neural networks. *ArXiv abs/1706.05098* (2017).

[118] RUDICK, R. A., LEE, J.-C., SIMON, J., AND FISHER, E. Significance of t2 lesions in multiple sclerosis: A 13-year longitudinal study. *Annals of Neurology 60*, 2 (Aug. 2006), 236–242.

[119] SALEM, M., VALVERDE, S., CABEZAS, M., PARETO, D., OLIVER, A., SALVI, J., ROVIRA, À., AND LLADÓ, X. A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical 25* (2020), 102149.

[120] SEPAHVAND, N., HASSNER, T., ARNOLD, D., AND ARBEL, T. *CNN Prediction of Future Disease Activity for Multiple Sclerosis Patients from Baseline MRI and Lesion Labels: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I*. 01 2019, pp. 57–69.

[121] SEPAHVAND, N. M., ARNOLD, D. L., AND ARBEL, T. Cnn detection of new and enlarging multiple sclerosis lesions from longitudinal mri using subtraction images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 127–130.

[122] SEPAHVAND, N. M., ARNOLD, D. L., AND ARBEL, T. Cnn detection of new and enlarging multiple sclerosis lesions from longitudinal mri using subtraction images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 127–130.

[123] SEPAHVAND, N. M., HASSNER, T., ARNOLD, D. L., AND ARBEL, T. CNN prediction of future disease activity for multiple sclerosis patients from baseline MRI

and lesion labels. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2019, pp. 57–69.

[124] SHARMA, D., PURUSHOTHAM, S., AND REDDY, C. K. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports 11*, 1 (2021), 1–18.

[125] SHINOZAKI, T., AND WATANABE, S. Structure discovery of deep neural network based on evolutionary algorithms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 4979–4983.

[126] SHORTEN, C., AND KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data 6*, 1 (July 2019).

[127] SINGH, R., BHARTI, V., PUROHIT, V., KUMAR, A., SINGH, A. K., AND SINGH, S. K. Metamed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognition 120* (2021), 108111.

[128] SLED, J., ZIJDENBOS, A., AND EVANS, A. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging 17*, 1 (1998), 87–97.

[129] SPASOV, S., PASSAMONTI, L., DUGGENTO, A., LIO, P., AND TOSCHI, N. A multimodal convolutional neural network framework for the prediction of alzheimer's disease. vol. 2018, pp. 1271–1274.

[130] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15* (06 2014), 1929–1958.

[131] SRIVASTAVA, Y., MURALI, V., DUBEY, S. R., AND MUKHERJEE, S. Visual question answering using deep learning: A survey and performance analysis. In *Communications in Computer and Information Science*. Springer Singapore, 2021, pp. 75–86.

[132] STAHLSCHMIDT, S. R., ULFENBORG, B., AND SYNNERGREN, J. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics 23*, 2 (Jan. 2022).

[133] STONNINGTON, C. M., CHU, C., KLÖPPEL, S., JACK, C. R., ASHBURNER, J., AND FRACKOWIAK, R. S. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *NeuroImage 51*, 4 (July 2010), 1405–1413.

[134] SUMMAIRA, J., LI, X., SHOIB, A. M., AND ABDUL, J. A review on methods and applications in multimodal deep learning. *arXiv preprint arXiv:2202.09195* (2022).

[135] THUNG, K.-H., YAP, P.-T., AND SHEN, D. Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017, pp. 160–168.

[136] TIAN, X., LIU, J., KUANG, H., SHENG, Y., WANG, J., AND INITIATIVE, T. A. D. N. Mri-based multi-task decoupling learning for alzheimer's disease detection and mmse score prediction: A multi-site validation, 2022.

[137] VANDENHENDE, S., GEORGOULIS, S., PROESMANS, M., DAI, D., AND GOOL, L. V. Revisiting multi-task learning in the deep learning era. *CoRR abs/2004.13379* (2020).

[138] WAHLUND, L. O., BARKHOF, F., FAZEKAS, F., BRONGE, L., AUGUSTIN, M., SJOGREN, M., WALLIN, A., ADER, H., LEYS, D., PANTONI, L., PASQUIER, F., ERKINJUNTTI, T., AND SCHELTENS, P. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke 32*, 6 (June 2001), 1318–1322.

[139] WANG, T.-C., LIU, M.-Y., TAO, A., LIU, G., KAUTZ, J., AND CATANZARO, B. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

[140] WANG, Y., AND YAO, Q. Few-shot learning: A survey. *CoRR abs/1904.05046* (2019).

[141] WANG, Y., YAO, Q., KWOK, J. T., AND NI, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. 53*, 3 (jun 2020).

[142] WANG, Z., DUAN, T., FANG, L., SUO, Q., AND GAO, M. Meta learning on a sequence of imbalanced domains with difficulty awareness. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA, USA, oct 2021), IEEE Computer Society, pp. 8927–8937.

[143] WILLEMINK, M. J., KOSZEK, W. A., HARDELL, C., WU, J., FLEISCHMANN, D., HARVEY, H., FOLIO, L. R., SUMMERS, R. M., RUBIN, D. L., AND LUNGREN, M. P. Preparing medical imaging data for machine learning. *Radiology 295*, 1 (Apr. 2020), 4–15.

[144] XIAO, Y., CODEVILLA, F., GURRAM, A., URFALIOGLU, O., AND LOPEZ, A. M. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems 23*, 1 (Jan. 2022), 537–547.

[145] YALA, A., LEHMAN, C., SCHUSTER, T., PORTNOI, T., AND BARZILAY, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology 292*, 1 (July 2019), 60–66.

[146] YAMASHITA, R., NISHIO, M., DO, R. K. G., AND TOGASHI, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging 9*, 4 (June 2018), 611–629.

[147] YAN, Y., RICCI, E., SUBRAMANIAN, R., LANZ, O., AND SEBE, N. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *2013 IEEE International Conference on Computer Vision* (2013), pp. 1177–1184.

[148] YANG, L., AND SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing 415* (Nov. 2020), 295–316.

[149] YANG, Y.-Y., CHOU, C.-N., AND CHAUDHURI, K. Understanding rare spurious correlations in neural networks. *arXiv preprint arXiv:2202.05189* (2022).

[150] YAO, H., WANG, Y., WEI, Y., ZHAO, P., MAHDAVI, M., LIAN, D., AND FINN, C. Meta-learning with an adaptive task scheduler, 2021.

[151] YAO, H., WU, X., TAO, Z., LI, Y., DING, B., LI, R., AND LI, Z. Automated relational meta-learning, 01 2020.

[152] YAP, J., YOLLAND, W., AND TSCHANDL, P. Multimodal skin lesion classification using deep learning. *Experimental Dermatology 27*, 11 (Sept. 2018), 1261–1267.

[153] YIM, J., JUNG, H., YOO, B., CHOI, C., PARK, D., AND KIM, J. Rotating your face using multi-task deep neural network. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 676–684.

[154] YU, J., ZHU, Z., WANG, Y., ZHANG, W., HU, Y., AND TAN, J. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition 108* (Dec. 2020), 107563.

[155] ZAHEER, R., AND SHAZIYA, H. A study of the optimization algorithms in deep learning. In *2019 Third International Conference on Inventive Systems and Control (ICISC)* (Jan. 2019), IEEE.

[156] ZANG, X., YAO, H., ZHENG, G., XU, N., XU, K., AND LI, Z. Metalight: Value-based meta-reinforcement learning for traffic signal control. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 01 (Apr. 2020), 1153–1160.

[157] ZECH, J. R., BADGELEY, M. A., LIU, M., COSTA, A. B., TITANO, J. J., AND OERMANN, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine 15*, 11 (Nov. 2018), e1002683.

[158] ZHANG, D., AND SHEN, D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage 59*, 2 (2012), 895–907.

[159] ZHANG, H., LIU, C., ZHANG, W., ZHENG, G., AND YU, Y. GeneraLight: Improving environment generalization of traffic signal control via meta reinforcement learning. In *Proceedings of the 29th ACM International Conference on Information &amp Knowledge Management* (Oct. 2020), ACM.

[160] ZHANG, P., LI, J., WANG, Y., AND PAN, J. Domain adaptation for medical image segmentation: A meta-learning method. *Journal of Imaging 7*, 2 (2021).

[161] ZHANG, W., DENG, L., ZHANG, L., AND WU, D. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica* (2022), 1–25.

[162] ZHANG, X. S., TANG, F., DODGE, H. H., ZHOU, J., AND WANG, F. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining* (New York, NY, USA, 2019), KDD '19, Association for Computing Machinery, p. 2487–2495.

[163] ZHANG, Y., SIDIBÉ, D., MOREL, O., AND MÉRIAUDEAU, F. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing 105* (2021), 104042.

[164] ZHANG, Y., AND YANG, Q. An overview of multi-task learning. *National Science Review 5* (01 2018), 30–43.

[165] ZHAO, S., BHARATI, R., BORCEA, C., AND CHEN, Y. Privacy-aware federated learning for page recommendation. In *2020 IEEE International Conference on Big Data (Big Data)* (2020), pp. 1071–1080.

[166] ZHENG, W., YAN, L., WANG, F.-Y., AND GOU, C. Learning from the guidance: Knowledge embedded meta-learning for medical visual question answering. In *Communications in Computer and Information Science*. Springer International Publishing, 2020, pp. 194–202.

[167] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H., AND HE, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE 109*, 1 (Jan. 2021), 43–76.

[168] ZOPH, B., AND LE, Q. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations* (2017).