Multimodal Guidance for Medical Image Classification

by

Mayur Mallya

B.Tech., National Institute of Technology Karnataka, India, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Mayur Mallya**
Degree:	**Master of Science**
Thesis title:	**Multimodal Guidance for Medical Image Classification**
Committee:	**Chair:** Yağiz Aksoy Assistant Professor, Computing Science **Ghassan Hamarneh** Supervisor Professor, Computing Science **Manolis Savva** Committee Member Assistant Professor, Computing Science **Ali Mahdavi-Amiri** Examiner Assistant Professor of Professional Practice, Computing Science

Abstract

Medical imaging is a cornerstone of therapy and diagnosis in modern medicine. However, the choice of imaging modality for a particular theranostic task typically involves trade-offs between the feasibility of using a particular modality (e.g., short wait times, low cost, fast acquisition, reduced radiation/invasiveness) and the expected performance on a clinical task (e.g., diagnostic accuracy, efficacy of treatment planning and guidance). The goal of this thesis is to examine the ability to apply the knowledge learned from the less feasible but better-performing (*superior*) modality to guide the utilization of the more-feasible yet under-performing (*inferior*) modality and steer it towards improved performance. To this end, we develop a lightweight guidance model – an autoencoder-like deep neural network – that learns a mapping from the latent representation of the inferior modality to the latent representation of its superior counterpart. With the incorporation of this model in the classification framework of the inferior modality, we aim to compensate for the absence of the superior modality during inference time. We focus on the application of deep learning for image-based diagnosis and examine the advantages of our method in the context of two clinical applications: multi-task skin lesion classification from clinical and dermoscopic images and brain tumor classification from multi-sequence magnetic resonance imaging (MRI) and histopathology images. For both these scenarios, we show a boost in the diagnostic performance of the inferior modality without requiring the superior modality. Furthermore, in the case of brain tumor classification, our method outperforms the model trained on the superior modality while producing comparable results to the model that uses both modalities during inference.

Keywords: Deep Learning; Multimodal Learning; Image Classification; Student-Teacher Learning; Brain Tumors; Skin Lesions

Dedication

I like to dedicate this work to the people who worked towards keeping us safe during the COVID-19 pandemic.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my senior supervisor, Prof. Ghassan Hamarneh, starting from giving me an opportunity to be a part of his wonderful lab to helping me publish my first research paper at a top conference. I am immensely grateful for his unwavering guidance, support, and patience throughout my relatively long Master's program. Secondly, I would like to thank the members of my examining committee, Prof. Manolis Savva, Prof. Ali Mahdavi-Amiri, and Prof. Yağiz Aksoy, for their valuable feedback and time on this thesis.

I would further like to thank my labmates at the Medical Image Analysis Lab for the insightful discussions and feedback on my research and for making working at the lab a more enjoyable experience. Special thanks to my good friend and labmate, Kumar Abhishek, for all the advice, assistance, and motivation throughout my MSc program. I am also grateful to Ben Cardoen and Weina Jin for their patience and mentorship in making me a better researcher.

I would like to acknowledge the Mitacs Globalink program for awarding me the Graduate Fellowship in 2019, which was instrumental in landing a research-based graduate program in Canada. I am also grateful for the Mitacs Globalink Research Internship Award in 2017, which exposed me to Canadian universities and made it easy to choose the location for my graduate studies. Last but not least, I would like to thank the Compute Canada Federation (presently, Digital Research Alliance of Canada) for providing the much-needed computational resources for running the experiments described in this thesis.

Finally, I would like to extend my eternal gratitude to my family members for their unbounded encouragement and belief in my life decisions and for making me what I am today.

Table of Contents

List of Tables

List of Figures

List of Acronyms

MRI	Magnetic Resonance Imaging/Image
WSI	Whole Slide Image
CNN	Convolutional Neural Network
FC	Fully Connected
MIL	Multiple Instance Learning
CLAM	Clustering-constrained Attention-based MIL
CE	Cross Entropy
MSE	Mean Squared Error
ML	Machine Learning
DL	Deep Learning
F1	Micro-averaged F1 score
BA	Balanced accuracy score
ROC	Receiver Operating Characteristics
AUROC	Area Under the ROC
S-T	Student-Teacher
KD	Knowledge Distillation
PET	Positron Emission Tomography
CT	Computed Tomography
H&E	Hematoxylin and Eosin

List of Notations

$\mathcal{I}$	Inferior modality
$\mathcal{S}$	Superior modality
$\mathcal{X}$	Input space
$\mathcal{Y}$	Label space
N	Training data size
x	Input image
z	Latent representation
y	Ground truth classification label
$\hat{y}$	Predicted classification label
θ	Model parameters
$C(\cdot)$	Classification model
$E(\cdot)$	Encoder of the classification model
$D(\cdot)$	Decoder of the classification model
$G(\cdot)$	Guidance model
$F(\cdot)$	Guided classification model
$\mathcal{L}(\cdot)$	Loss function
w	Loss weights
W	WSI
a	Attention score
$A(\cdot)$	Attention network
$\mathbb{R}$	Set of real numbers
S	7-point score
$\hat{y}_{7pt}$	Predicted melanoma inference label

Chapter 1

Introduction

1.1 Motivation

Multimodal learning aims at analyzing the heterogeneous data in the same way animals perceive the world – by a holistic understanding of the information gathered from all the sensory inputs. The complementary nature of this multi-sensor input enables the animal in better navigating the surroundings than a single sensory input. Different sensory inputs provide different characteristics of the same object, all of which together provide a comprehensive overview of the particular object. The different ways in which the animal experiences the object, such as the sight, sound, odour, etc., in simple terms, encapsulates the word '*modality*'.

Multimodal learning, or multimodal representation learning, a sub-field of machine learning (ML), has seen a massive surge of interest in the recent years [1, 2, 3, 4, 5]. This can primarily be attributed to two reasons. Firstly, the availability of large volumes of multimodal data, owing to the ubiquity of a wide range of digital sensors that capture different types of signals at different resolutions. Secondly, the emergence of deep learning (DL) algorithms [6] that thrive on large volumes of data. The powerful feature abstraction abilities of DL models coupled with the automatic selection of relevant features has motivated the success of multimodal learning.

Since multimodal data are more informative than unimodal data, in a clinical setting, where there is a high cost on the decisions that are dependent on the available data, it is of utmost importance to gather the maximum amount of information before finalizing the diagnosis and treatment decisions. A comprehensive assessment of a patient's health is crucial for a clinician to make an informed disease diagnosis and formulate an accurate management strategy. Typically, this step involves the acquisition of complementary biomedical data from the target organ of the patient across multiple different modalities. For instance, the simultaneous acquisition of functional and anatomical imaging data is a

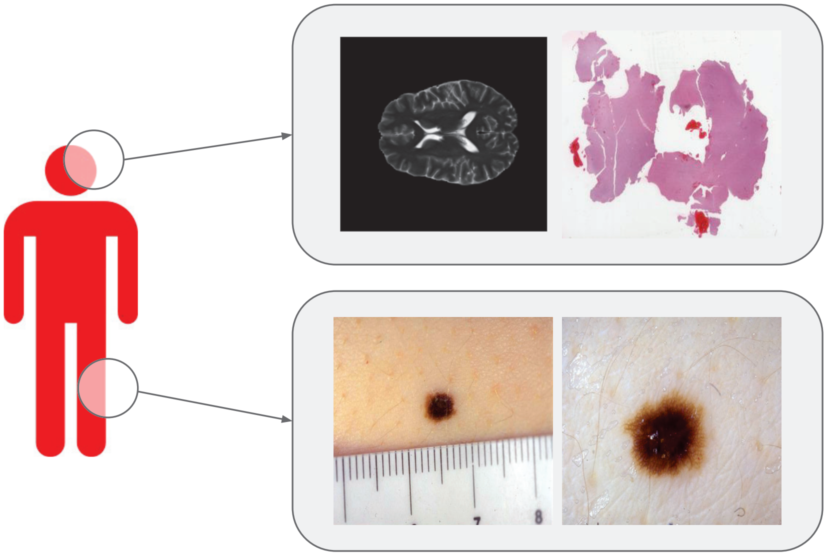


Figure 1.1: Multimodal medical images corresponding to a brain tumor (*top*) and a skin lesion (*bottom*) of a patient. The brain tumor is imaged using an MRI scan of the entire brain (*left*) and a microscopy visualization of the tumorous tissue sections of the brain (*right*). The skin lesion is imaged using clinical (*left*) and dermoscopy (*right*) imaging techniques. Sources of images are as follows: patient [13], brain MRI and tissue [14], skin lesions [15].

common practice in the modern clinical setting. Two of the popular practices are the simultaneous acquisition of Positron Emission Tomography (PET) scans alongside Magnetic Resonance Imaging (MRI) [7] or Computed Tomography (CT) [8] scans. While the PET scans provide accurate quantitative information on the metabolic activity of the target organ, the MRI or CT scans provide the anatomical details about the spatial arrangement of the body organs and tissue. The simultaneous acquisition, when accurately aligned, significantly improves the identification and localization of the diseases. In the same way, cancer diagnosis and prognosis decisions are increasingly a result of a thorough examination of both the genotypic and phenotypic modalities [9, 10]. Here, the phenotypic modalities, such as the histologic tissue slides, provide the spatial and morphological information about the tumor while the genomic analysis of the tissues provides the quantitative genetic mutations and aberrations caused by the tumors. Together, the two modalities not only aid in the better identification of the cancer subtype but also improve our understanding of cancer as a whole [11, 12].

Although the complimentary use of multiple modalities can improve the clinical diagnosis, the acquisition of different modalities is not equally feasible. The modalities that are easy to acquire are typically less informative and are to be supplemented subsequently by the more informative modalities. In a practical scenario, the modalities that are easy to acquire are typically used for the preliminary diagnosis of the disease. The modalities that provide critical information about the disease, on the other hand, are less feasible to acquire due to a multitude of reasons. The most common reasons are the long wait times, higher cost of scan, slower acquisition, higher radiation exposure, and/or invasiveness. This leads to a trade-off between the feasibility of acquisition of a modality and its the expected performance on a clinical task.

An imaging system that provides a confident diagnosis may be prohibitively expensive or fraught with long wait times; a higher anatomical or functional resolution imaging may only be possible with a modality that involves invasive surgical procedures or ionizing radiations. For instance, a CT scan provides a precise anatomical structure of the body organs when compared to a simple X-ray scan, mainly due to the powerful doses of radiation. One CT scan can have the same dosage of radiation as 200 X-ray scans [16]. Although the benefits outweigh the risks, repeated exposure to CT scans, especially in children, is discouraged due to the associated links to cancerous mutations [17, 18]. Another instance of such a trade-off is the acquisition of the high-resolution histologic tissue images that provide rich cellular information necessary for tumor identification. While highly informative, these images come at the cost of an expensive, time-consuming, and invasive surgical procedure with associated risks of patient bleeding and infections [19]. Unsurprisingly, in most cases, it is the expensive modality that provides the critical piece of information for diagnosis.

For simplicity, hereinafter, we refer to the over-performing modality with the less-feasible acquisition as the *superior* modality, and the more-feasible but under-performing one as the *inferior*. We note that a particular modality may be regarded as inferior in one context and superior in another. For example, MRI is superior to ultrasound images for delineating cancerous lesions but inferior to histologic images in deciding cancer subtype or grade.

Given the existence of the aforementioned trade-off, it would be advantageous to leverage the inferior modalities in order to alleviate the need for the acquisition of the superior modality. However, this is reasonable only when the former can be as informative as the latter. To this end, we propose a novel DL-based method that leverages existing datasets of paired inferior and superior modalities during a training phase in order to enhance the diagnosis performance achievable by only the inferior modality during the inference phase. The practical clinical equivalent of the scenario is to leverage the typically abundant, multi-modal data of the previous patients in order to improve the diagnosis of the current patient using only the inferior unimodal data.

We make use of the concepts from the prior works on multimodal and student-teacher learning to effectively transfer the knowledge from the superior modality to the inferior. In order to test the efficacy of the proposed method, we conduct experiments on two disparate multimodal datasets of vastly different resolutions: a multi-task skin lesion classification dataset consisting of clinical and dermoscopic images and a brain tumor classification dataset consisting of multi-sequence MRI and histopathology images. Our experiments on these datasets across several classification tasks demonstrate the validity and utility of the proposed method. We summarize the contributions of this thesis in section 1.4.

1.2 Brain Tumors

A brain tumor is one of the common diseases of the central nervous system (CNS) and is a complex and fatal disease. A statistical report published by Ostrom *et al.* in 2018 [20] based on the survey conducted in the United States between 2011 and 2015 on the primary brain and other CNS tumors, found an overall age-adjusted incidence rate of 23.03 per 100,000 population. Although the incidence rate of brain cancer compared to cancers of other organs is low [21], they are associated with significant mortality and morbidity. The five- and ten-year relative survival rates for patients with malignant brain and other CNS tumors were estimated to be 35.0% and 29.3% respectively [20]. The most common type of malignant brain tumor in adults is glioma, and it corresponds to 81% of all malignant brain and CNS tumors. Glioblastoma, the most aggressive subtype of glioma, accounts for ∼45% of glioma cases and has a five-year overall survival rate as low as ∼5% [22].

The diagnosis, grading, and stratification of different subtypes of brain tumors are conventionally done by pathologists who, using a microscope, examine the processed and stained tissue specimens fixed on glass slides. This technique is considered the golden standard for clinical diagnosis. In 2016, when the World Health Organization revised the CNS tumor classification criteria to include the molecular parameters [11], the whole slide tissue analysis remained the most informative visual modality for tumor diagnosis. However, non-invasive and relatively safe MRIs are routinely used for preliminary diagnosis in the clinical setting. While MRI is useful in locating the tumorous regions within an organ, due to the complexity and heterogeneity of the different tumor subtypes, MRI alone is not sufficient and is typically complemented by the more informative pathology slides for an accurate diagnosis and the subsequent treatment planning.

Although the pathology-based tissue slides are more informative than MRI, the acquisition of the former is comparatively less feasible than the latter. Figure 1.2 shows the steps involved in the acquisition and the subsequent digitization of the pathology slides [25]. The first step involves the collection of sufficient good-quality tissue for diagnosis. This is done

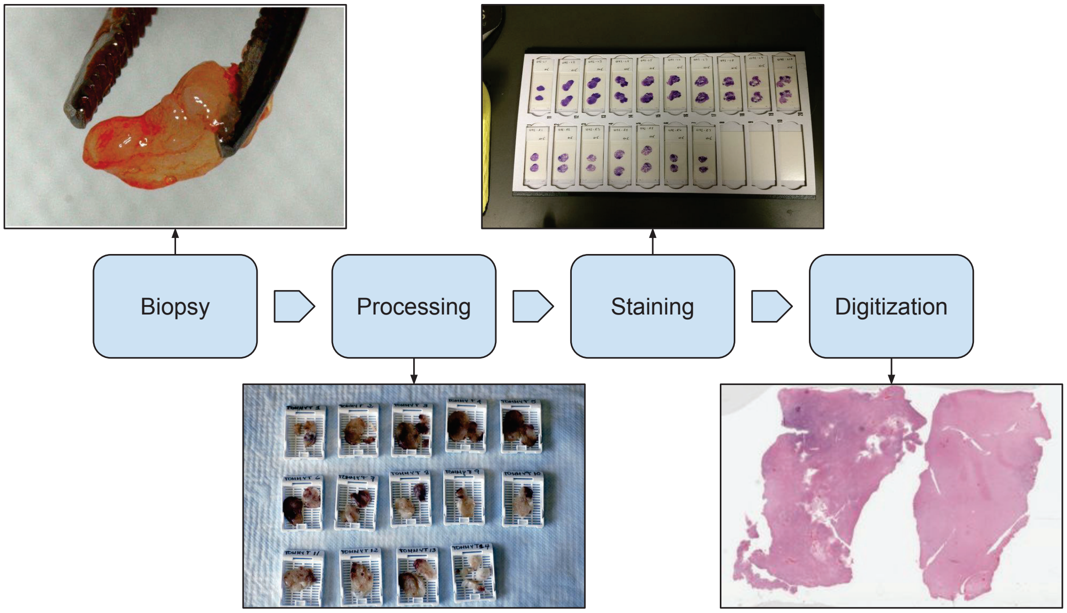


Figure 1.2: Steps involved in the acquisition of whole slide images of the tumorous tissue sections. Sources of different images are as follows: biopsy [23], processing and staining [24], digitization [14].

via biopsy, an invasive surgical procedure that comes with the risk of bleeding and infections [19]. The tissues are then immersed into a fixative solution to prevent the breaking down of tissues and are further embedded into cassette-sized paraffin boxes to minimize the changes in the shape of the tissue. In the third step, the tissues, which are semi-transparent at this stage, are stained using chemical agents called dyes in order to make them visible under the microscope. Finally, the stained tissues are imaged using high-resolution microscopes to produce digitized whole slide images (WSI).

On the other hand, the acquisition of MRIs involves a non-invasive procedure as shown in Figure 1.3. MRI scanners use non-ionizing radio waves to excite the water molecules in the brain. The strong magnetic field created by the MRI scanner enables the detection of the excited water molecules and in turn, aids in the imaging of the brain [26]. By varying the pulse sequence of the radio waves, multiple complementary images with different contrasts can be generated. Figure 1.3 shows the four common sequences- native (T1), T2-weighted (T2), post-contrast T1-weighted (T1Gd), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) (*top to bottom*).

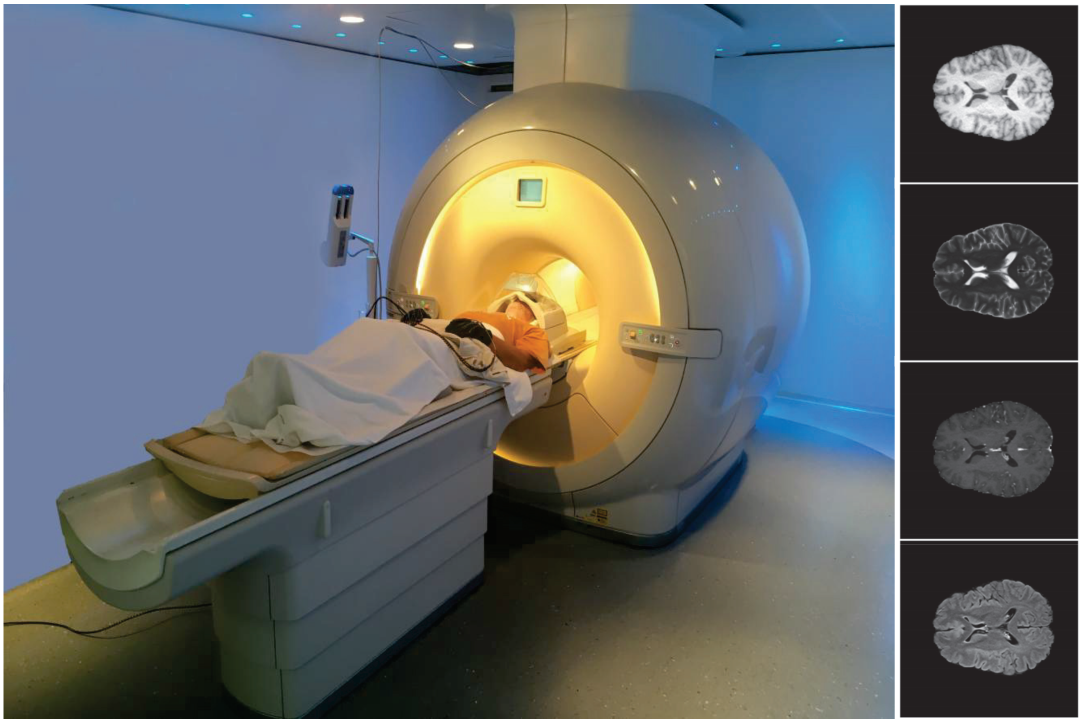


Figure 1.3: Acquisition of MRI scans of the patient's brain (*left*) and the acquired multi-sequence brain MRI (*right*). The sequences include native (T1), T2-weighted (T2), post-contrast T1-weighted (T1Gd), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) (*top to bottom*). Sources of images are as follows: acquisition [27], multi-sequence MRI [14].

1.3 Skin Cancer

Skin cancer is one of the most common types of cancer with over 1 million new cases and 100,000 fatalities in 2018 alone [21]. The incidence of skin lesions is more common in populations with fairer skin complexions and is observed to have an increase in incidence with decreasing latitude – the highest annual incidence rates recorded in Australia [28]. A majority of cases are caused due to high and intermittent exposure to ultraviolet radiation from the Sun. Melanoma, the most aggressive form of skin cancer, despite its low prevalence compared to other forms, is a major cause of death from skin cancer [28]. The American Cancer Society estimates about 7,650 fatalities due to Melanoma in the United States this year [29]. However, early detection has shown to significantly improve patient outcomes [30], making early diagnosis and treatment extremely important.

While a skin biopsy and the subsequent examination of the microscopic tissues is the golden standard for skin lesion diagnosis, dermatologists have long relied on the visual inspection of the skin lesions. Dermoscopy, also known as epiluminescence microscopy, is the

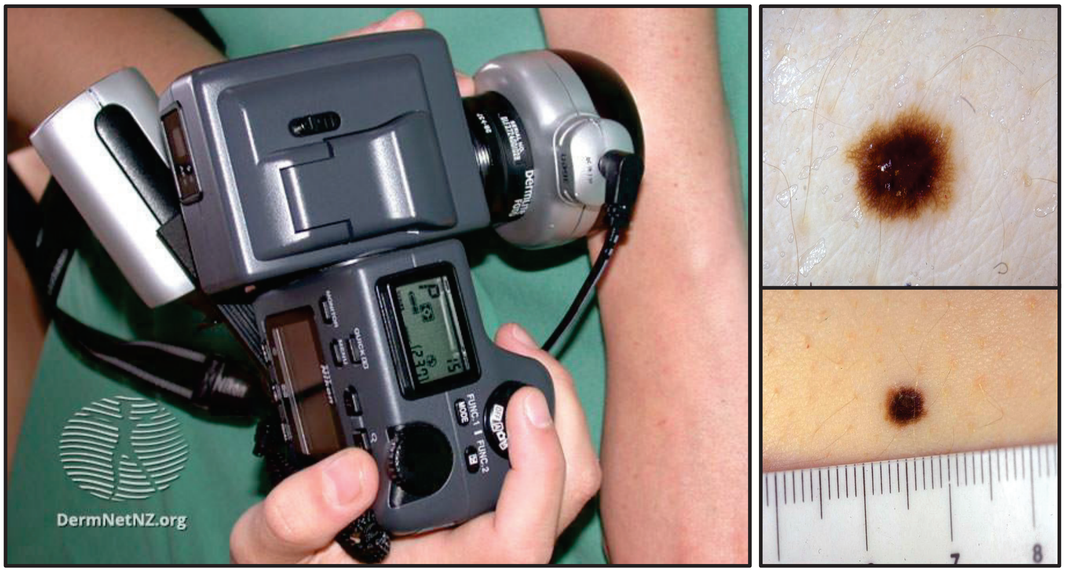


Figure 1.4: Acquisition of dermoscopic image of a skin lesion (*left*) and the acquired dermoscopic image (*top-right*). The clinical image of the corresponding skin lesion (*bottom-right*). Sources of images are as follows: acquisition [33], skin lesions [15].

most popular non-invasive imaging technique that provides detailed morphological and visual properties of the subsurface skin structures of the pigmented lesions that are not visible to the unaided eye [31]. The use of dermoscopy has been shown to improve the diagnostic performance of experienced dermatologists, with improvements as high as 49% in the accuracy of melanoma diagnosis [32].

The acquisition of dermoscopic images of skin lesions requires a dermoscope (or dermatoscope), a hand-held device with a high-quality magnifying lens, and a powerful lighting system. Figure 1.4 shows the acquisition of the image of a skin lesion using a dermoscope (*left*). On the right side of Figure 1.4 is an example of the dermoscopic image (*top*) and the corresponding clinical image (*bottom*). The acquisition of clinical images, on the other hand, is relatively convenient as they are captured using ubiquitous and inexpensive mobile cameras. This also means the acquired clinical images are relatively less standardized and include artefacts such as rulers, as seen in the clinical image in Figure 1.4.

1.4 Thesis Contributions

The superior modality is more informative than the inferior modality with respect to disease diagnosis but the acquisition of the superior modality is less feasible than the inferior modality. The goal of this thesis is to address this trade-off between any two modalities that exhibit this behavior. The key contributions of this thesis are as follows:

- While previous works on multimodal disease classification rely on the presence of both modalities during the inference time, our proposed method learns from the multimodal data during training and uses only the unimodal data from the inferior modality during the inference.
- We propose a novel method based on the multimodal and student-teacher learning frameworks to effectively transfer the knowledge from the superior modality to the inferior. Our proposed guidance model is a lightweight model that is easy to implement and adapt even for modalities of different dimensions and scales.
- Our experiments on two multimodal datasets, corresponding to diseases of two different body organs (brain tumors and skin lesions), demonstrate that the proposed method improves the diagnostic performance achievable by the inferior modality. Additionally, in the case of brain tumor classification, our proposed method using the inferior modality outperformed diagnostic performance achievable by the superior modality and was comparable to that of the multimodal diagnosis using both modalities.

This thesis is based on the work that has been accepted to the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022) [34]. A significant portion of the text in this thesis is copied verbatim from the MICCAI paper. The preprint is made available on arXiv under the title: Deep Multimodal Guidance for Medical Image Classification [35]. The codes and the pre-trained models would be made publicly available at `https://github.com/mayurmallya/DeepGuide`.

> **Mayur Mallya**, Ghassan Hamarneh, **Deep Multimodal Guidance for Medical Image Classification**, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022.

1.5 Thesis Outline

The rest of the thesis is organized as follows: Chapter 2 provides a detailed list of the related works that are closely related to the methodology proposed in the thesis. Chapter 3 presents the mathematical formulation of the problem statement and the proposed methodology. In Chapter 4, we describe the experimental setup by illustrating the datasets, preprocessing strategies, and hyperparameters used in the experiments. The results of the different experiments are discussed in Chapter 5. Finally, the conclusion and the future works are presented in Chapter 6.

Chapter 2

Related Work

2.1 Multimodal learning

Multimodal learning, in general, aims at the utilization of ubiquitous multimodal data to learn comprehensive latent space representations that can be used for a variety of downstream tasks. Despite the recent surge of interest in multimodal learning [1, 2, 3, 4, 5], one of the commonly encountered challenges in the field is the heterogeneity gap between the different modalities. Specifically, this corresponds to the situation where the latent space of modalities A and B differ semantically even though the two modalities capture the same object. Thus, narrowing the heterogeneity gap in order to facilitate learning of a common latent space for multiple modalities becomes the primary objective of multimodal learning.

To this end, based on how different approaches narrow the heterogeneity gap, Guo *et al.* [2] categorize the deep multimodal learning methods into three types of frameworks – joint representation, coordinated representation, and translation frameworks. Joint representation projects the unimodal representations into a common latent space where they are fused together to form a multimodal representation. Coordinated representation aims to learn modality-specific representations that are constrained to each other (usually by a loss function). Finally, the translation frameworks, as the name suggests, involve the translation from one modality to another and aim to learn the latent space between the source and target modalities. Figure 2.1 presents an overview of the different frameworks proposed by Guo *et al.* [2].

Joint representation framework

The fusion of heterogeneous latent representations from different modalities to form a common multimodal representation has shown to improve the performance over unimodal representations across a variety of tasks such as image classification [36, 37, 38, 39, 15], segmentation [4, 40, 41], video event recognition [42, 43, 44], etc. The joint representation framework is used in cases where the multimodal data is available both at the training

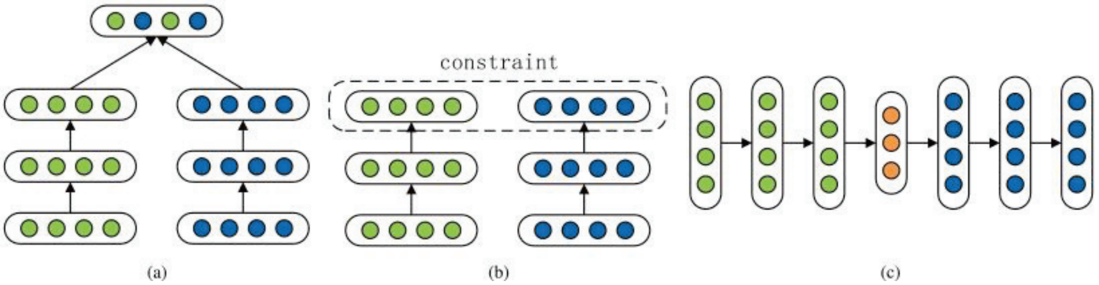


Figure 2.1: Multimodal learning frameworks – (a) joint representation framework, (b) coordinated representation framework, and (c) translation framework. The figure is copied from the work of Guo *et al.* [2].

and the inference time. The focus of research is the optimal fusion strategy that aims to answer when and how to efficiently fuse the supposedly heterogeneous and redundant features. *When to fuse?* While the registered multimodal image pairs allow for an input-level data fusion [41, 45], a majority of works rely on feature-level fusion not only due to the dimensionality mismatch at the input but also for the flexibility of fusion it offers [38, 39, 15]. Additionally, some works make use of a decision-level fusion framework that leverages the ensemble learning strategies [46]. *How to fuse?* The most popular fusion strategy to date is the straightforward concatenation of the extracted features [47, 15]. However, recent works aim to learn the interactions across multimodal features using strategies like the Kronecker product in order to model pairwise feature interactions [38]. Kawahara *et al.* [15] concatenate the latent representations of the clinical and dermoscopic skin lesions at the feature level. While the winners of the RadPath 2020 challenge [48] follow the same strategy, the winners of the RadPath 2019 challenge [49] use the decision-level fusion of the modality-wise predictions.

Coordinated representation framework

Unlike the joint representation framework where the modality-specific latent representations are fused to form a multimodal representation, in the coordinated representation framework, separate representations are learned for each modality but are coordinated via constraints (Figure 2.1(b)). In general, the constraints encourage (or discourage) a specific characteristic such as the similarity, correlation, or orthogonality between the modality-specific latent representations as a way to mitigate the heterogeneity. Such constraint-based representation learning methods are commonly seen at the intersection of vision and language [1, 2]. Frome *et al.* [50] showed that leveraging the semantic similarity of the textual data can improve image-based object identification. Conversely, Kiros *et al.* [51] and Pan *et al.* [52] improved respectively the image and video captioning by coordinating the latent space between the images/videos and sentences. Coordinated representations are also found useful for cross-modal retrieval tasks [53, 54, 55] as this framework allows for the coordinated representations of each modality to be used individually for downstream tasks during inference.

Baltrušaitis *et al.* [1] coined the term *multimodal co-learning* specifically for methods that aim to assist the analysis of a resource-poor modality by exploiting the knowledge of the resource-rich modality. The resources can refer to the quality of the input data, the availability of modality-specific annotations, the presence of noise-free labels, and so on. The resource-rich modality is used only during the model training and not during the inference. Ngiam *et al.* [56] used a bimodal autoencoder to learn multimodal representations for paired audio and video signals. The learned representations were used for unimodal downstream tasks of both audio and video inputs. Similarly, Moon *et al.* [57] use latent space transfer learning from audio to the video modality to improve the inference time lip-reading using only the video signal without accessing the audio signal.

Translation framework

The goal of this framework is to translate an element from one modality to the corresponding element in the other modality. The translation framework has been successful in a variety of tasks including machine translation [58, 59], text-to-image and text-to-speech generation [60, 61], image-to-image translation [62, 63, 64], and so on. Image-to-image translation has also been applied successfully in the medical domain [65], the most common applications being the inter-modality translation between the registered imaging modalities such as MRI, CT, and PET scans [66, 67]. Additionally, intra-modality translation has also been applied for tasks such as image denoising [66], generating the missing sequence [68], artefact correction [66], etc.

Although great success has been witnessed in image-to-image translation, the application of this framework in several medical applications is complicated or non-ideal due to the difference in the dimensionality (e.g. 2D to 3D) and size (e.g. millions to billions of voxels) between source and target modalities. The smaller size of the datasets also hinders the performance of the translation. Moreover, in most cases, the translation is not optimized for the subsequent downstream tasks which make it challenging for applications such as ours, where the end goal is to improve the disease diagnosis as opposed to modality translation.

2.2 Student-Teacher (S-T) learning

Often referred to as Knowledge Distillation (KD), S-T learning aims to transfer the knowledge learned from one model to another. The primary applications of S-T learning have so far been towards model compression and knowledge transfer [69]. Model compression aims

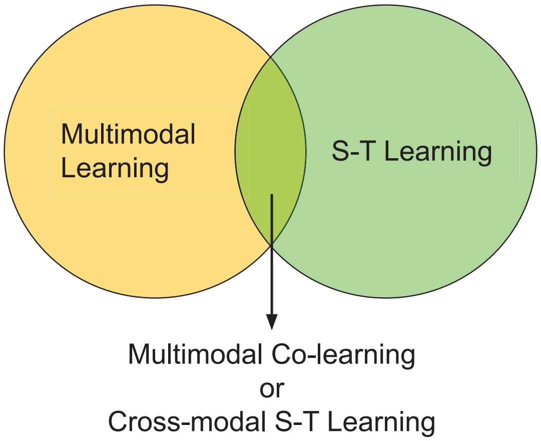


Figure 2.2: Overview of the related areas. Our proposed method places at the intersection of multimodal learning and S-T learning, and falls under the sub-fields of multimodal co-learning or cross-modal S-T learning.

at obtaining lightweight versions of deep networks while knowledge transfer is used in cases where there is a lack of labeled data.

Depending on how the knowledge is distilled from the teacher to the student, S-T learning can be of two types – logit learning and feature learning methods. While the logit learning methods such as the works proposed by Zhang *et al.* [70] and Wen *et al.* [71] use the soft predictions of the models as the transferable knowledge, feature learning methods proposed by Heo *et al.* [72], Huang *et al.* [73], and Tung *et al.* [74] use the intermediate latent representations to transfer knowledge from the teacher to the student model. Due to their flexibility in a DL setup, the feature learning methods, in most cases, have been observed to be more effective for knowledge transfer [69].

Cross-modal distillation is a type of KD where the teacher and student models contain modality-specific representations, typically with a better-performing teacher. Similar to other S-T learning methods, the goal of this framework is to leverage the latent representation of the teacher, which in this case is modality-specific, and distill the knowledge onto the student. While most such applications focus on KD across synchronized visual and audio data [75, 76, 77], KD methods for cross-modal medical image analysis mainly focus on segmentation [78, 79, 80]. These works make use of the pixel/voxel registered modalities, which enables them to leverage the anatomical structures that are more evident in the teacher modality to guide the student modality. However, with respect to the unregistered cross-modal KD, Sonsbeek *et al.* [81] recently proposed a multimodal KD framework for classifying chest x-ray images with the language-based electronic health records as the teacher and X-Ray images as the student network. Their work uses a probabilistic framework where the latent representations of the image-based x-ray modality are conditioned on the latent representations of the language-based electronic health record modality. Similar to the cross-modal KD works on segmentation, their method encourages the student only to mimic the latent distribution of the teacher modality, without adding its own knowledge to the final prediction – a drawback that we circumvent in our methodology.

To summarize, while prior works use multimodal medical images as input during inference to improve performance, our contribution is that we leverage multimodal data during training in order to enhance inference performance with only unimodal input. Figure 2.2 shows the placement of the proposed method among the related areas of multimodal and ST learning.

Chapter 3

Method

3.1 Problem formulation

Given a training dataset $\mathcal{X}$ of N paired images from inferior and superior modalities with corresponding ground truth target labels $\mathcal{Y}$, our goal is to learn a function F that maps novel examples of the inferior type to the corresponding target labels. Specifically, $\mathcal{X} = \{\mathcal{X}_\mathcal{I}, \mathcal{X}_\mathcal{S}\}$, with $\mathcal{X}_I = \{x^i_\mathcal{I}\}^N_{i=1}$ and $\mathcal{X}_S = \{x^i_\mathcal{S}\}^N_{i=1}$, is the set of paired inferior and superior images, i.e. $(x^i_\mathcal{I}, x^i_\mathcal{S})$ is the i-th pair. The set of training labels is $\mathcal{Y} = \{y_i\}^N_{i=1}$, where $y_i \in \mathcal{L}$ and $\mathcal{L} = \{l_1, l_2, ..., l_K\}$ is the label space representing the set of all K possible class labels (e.g., disease diagnoses). We represent F using a deep model with parameters θ, i.e., $\hat{y} = F(x_\mathcal{I}; \theta)$, where $\hat{y}$ is the model prediction.

3.2 Model optimization

The model F must leverage the paired multimodal data during the training and be able to predict the label using only the unimodal input (inferior modality) during the inference. The proposed method comprises 3 steps:

1. Train classifiers $C_\mathcal{I}$ and $C_\mathcal{S}$ that each independently predicts the target label y from $x_\mathcal{I}$ and $x_\mathcal{S}$ respectively.

2. Train a guidance model G to learn the mapping from the latent representation in $C_\mathcal{I}$ to that in $C_\mathcal{S}$.

3. Construct F that, first, maps $x_\mathcal{I}$ to the latent representation in $C_\mathcal{I}$, then using G maps that representation to the latent representation in $C_\mathcal{S}$ to perform the classification.

We now describe these steps in detail.

3.2.1 Classifiers $C_\mathcal{I}$ and $C_\mathcal{S}$

Given image pairs $(x^i_\mathcal{I}, x^i_\mathcal{S})$ and ground-truth labels y^i, we train two independent classification models $C_\mathcal{I}$ and $C_\mathcal{S}$ on the same task. Classifier $C_\mathcal{I}$ is trained to classify images of the

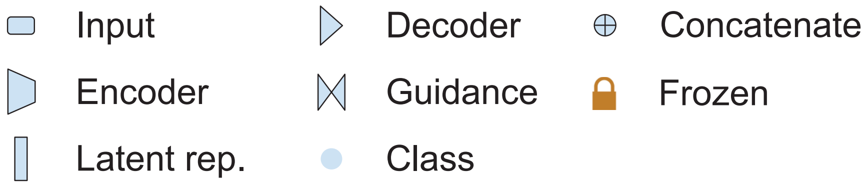


Figure 3.1: Notations used to denote the different components of the model architecture throughout the proposed methodology.

inferior modality $x_{\mathcal{I}}$, whereas $C_{\mathcal{S}}$ is trained to classify images of the superior modality $x_{\mathcal{S}}$. Denoting the predictions made by the two networks as $\hat{y}^i_{\mathcal{I}}$ and $\hat{y}^i_{\mathcal{S}}$, we have:

$$\begin{aligned} \hat{y}^i_{\mathcal{I}} &= C_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta_{\mathcal{I}}), \\ \hat{y}^i_{\mathcal{S}} &= C_{\mathcal{S}}(x^i_{\mathcal{S}}\ ;\ \theta_{\mathcal{S}}), \end{aligned} \tag{3.1}$$

where $C_{\mathcal{I}}$, and similarly $C_{\mathcal{S}}$, comprises an encoder E, which encodes the high-dimensional input image into a compact low-dimensional latent representation, and a decoder D, which decodes the latent representation by mapping it to one of the labels in $\mathcal{L}$. Denoting the encoder and decoder in $C_{\mathcal{I}}$ as $E_{\mathcal{I}}$ and $D_{\mathcal{I}}$, respectively, and similarly $E_{\mathcal{S}}$ and $D_{\mathcal{S}}$ in $C_{\mathcal{S}}$, we obtain:

$$\begin{aligned} \hat{y}^i_{\mathcal{I}} &= D_{\mathcal{I}} \circ E_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta_{E_{\mathcal{I}}}), \\ \hat{y}^i_{\mathcal{S}} &= D_{\mathcal{S}} \circ E_{\mathcal{S}}(x^i_{\mathcal{S}}\ ;\ \theta_{E_{\mathcal{S}}}), \end{aligned} \tag{3.2}$$

where $\circ$ denotes function composition, and $\theta_{E_{\mathcal{I}}}$ and $\theta_{E_{\mathcal{S}}}$ are the encoder parameters of $C_{\mathcal{I}}$ and $C_{\mathcal{S}}$ respectively. The encoders produce the latent representations z, i.e.:

$$\begin{aligned} z^i_{\mathcal{I}} &= E_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta_{E_{\mathcal{I}}}), \\ z^i_{\mathcal{S}} &= E_{\mathcal{S}}(x^i_{\mathcal{S}}\ ;\ \theta_{E_{\mathcal{S}}}). \end{aligned} \tag{3.3}$$

Latent codes $z^i_{\mathcal{I}}$ and $z^i_{\mathcal{S}}$ are inputs to corresponding decoders $D_{\mathcal{I}}$ and $D_{\mathcal{S}}$. Finally, $D_{\mathcal{I}}(z^i_{\mathcal{I}}; \theta_{D_{\mathcal{I}}})$ and $D_{\mathcal{S}}(z^i_{\mathcal{S}}\ ;\ \theta_{D_{\mathcal{S}}})$ yield predictions $\hat{y}^i_{\mathcal{I}}$ and $\hat{y}^i_{\mathcal{S}}$, respectively. Here, $\theta_{\mathcal{I}} = \{\theta_{E_{\mathcal{I}}}, \theta_{D_{\mathcal{I}}}\}$ and $\theta_{\mathcal{S}} = \{\theta_{E_{\mathcal{S}}}, \theta_{D_{\mathcal{S}}}\}$.

We use the cross-entropy loss (Equation 3.10) to optimize the classification models. We optimize our classification models $C_{\mathcal{I}}(\cdot)$ and $C_{\mathcal{S}}(\cdot)$ as:

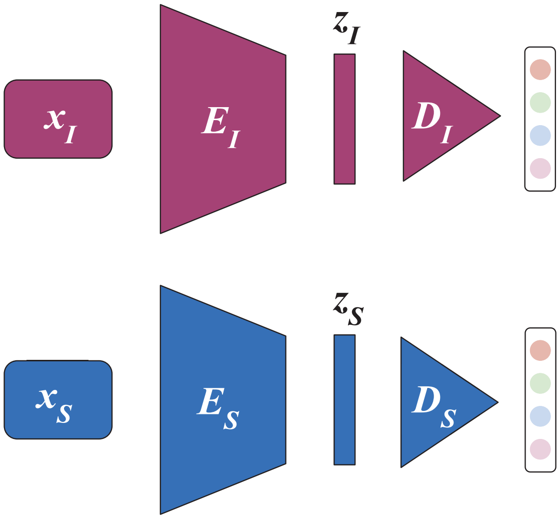


Figure 3.2: Two independent modality-specific classifiers $C_{\mathcal{I}}(\cdot)$ (*top*) and $C_{\mathcal{S}}(\cdot)$ (*bottom*) are trained to classify the images respectively from the inferior ($\mathcal{X}_{\mathcal{I}}$) and the superior ($\mathcal{X}_{\mathcal{S}}$) modality. Each classifier consists of a modality-specific encoder (denoted by E) that extracts features from the high-dimensional input image (x) to produce a lower-dimension latent representation (z). The decoder (D) predicts the class based on the latent representation.

$$\begin{aligned} \theta^*_{\mathcal{I}} &= \arg\min_{\theta_{\mathcal{I}}} \mathcal{L}_{CE}(C_{\mathcal{I}}(\mathcal{X}_{\mathcal{I}}\,;\,\theta_{\mathcal{I}}),\,\mathcal{Y}), \\ \theta^*_{\mathcal{S}} &= \arg\min_{\theta_{\mathcal{S}}} \mathcal{L}_{CE}(C_{\mathcal{S}}(\mathcal{X}_{\mathcal{S}}\,;\,\theta_{\mathcal{S}}),\,\mathcal{Y}). \end{aligned} \tag{3.4}$$

The notations of the different components of the architecture are provided in Figure 3.1. The overview of this step is provided in Figure 3.2. Algorithm 1 presents the PyTorch-style pseudocode corresponding to this step.

3.2.2 Guidance Model G

Guidance model, G is trained to map the latent representation of the inferior image, $z^i_{\mathcal{I}}$ to the latent representation of the paired superior image, $z^i_{\mathcal{S}}$. Denoting the estimated latent code as $\hat{z}^i_{\mathcal{S}}$ and the parameters of G as θ_G, we obtain:

$$\hat{z}^i_{\mathcal{S}} = G(z^i_{\mathcal{I}}\,;\,\theta_G). \tag{3.5}$$

Figure 3.3 presents an overview of the guidance model. We use the mean squared error loss (Equation 3.11) to encourage the similarity between the predicted and superior modality latent representations. We optimize the guidance model, $G(\cdot)$ as:

Algorithm 1 PyTorch-style pseudocode for training modality-specific classifiers $C_{\mathcal{I}}(\cdot)$ and $C_{\mathcal{S}}(\cdot)$.

Input : A multimodal dataset $\mathcal{X} = \{\mathcal{X}_{\mathcal{I}}, \mathcal{X}_{\mathcal{S}}\}$ with labels $\mathcal{Y}$. Specifically, $(\{x^i_{\mathcal{I}}, x^i_{\mathcal{S}}\},\ y^i)$ $\forall i \leq N$, where N is the size of the dataset.

Output: Learned modality-specific classifiers $C_{\mathcal{I}}(\cdot)$ and $C_{\mathcal{S}}(\cdot)$ with respective encoders $E_{\mathcal{I}}(\cdot)$ and $E_{\mathcal{S}}(\cdot)$ and decoders $D_{\mathcal{I}}(\cdot)$ and $D_{\mathcal{S}}(\cdot)$ with learned parameters $\theta^*_{E_{\mathcal{I}}}$, $\theta^*_{E_{\mathcal{S}}}$, $\theta^*_{D_{\mathcal{I}}}$, and $\theta^*_{D_{\mathcal{S}}}$ respectively.

```
while not converged do
    for $i = 1$ to $N$ do
        ▷ Model predictions
        $\hat{y}^i_{\mathcal{I}} = C_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta_{\mathcal{I}})$        ▷ $\theta_{\mathcal{I}} = \{\theta_{E_{\mathcal{I}}}, \theta_{D_{\mathcal{I}}}\}$
        $\hat{y}^i_{\mathcal{S}} = C_{\mathcal{S}}(x^i_{\mathcal{S}}\ ;\ \theta_{\mathcal{S}})$        ▷ $\theta_{\mathcal{S}} = \{\theta_{E_{\mathcal{S}}}, \theta_{D_{\mathcal{S}}}\}$
        ▷ Loss computation
        $\mathcal{L}_{\mathcal{I}} = \mathcal{L}_{CE}(\hat{y}^i_{\mathcal{I}}\ , y^i)$
        $\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{CE}(\hat{y}^i_{\mathcal{S}}\ , y^i)$
        ▷ Optimization ($\eta$: learning rate, $\nabla$: gradient)
        $\mathcal{L}_{\mathcal{I}}$.backward()        ▷ $\theta^*_{\mathcal{I}} \leftarrow \theta_{\mathcal{I}} - \eta\ \nabla\theta_{\mathcal{I}}$
        $\mathcal{L}_{\mathcal{S}}$.backward()        ▷ $\theta^*_{\mathcal{S}} \leftarrow \theta_{\mathcal{S}} - \eta\ \nabla\theta_{\mathcal{S}}$
    end for
end while
```

$$\theta^*_G = \arg\min_{\theta_G} \mathcal{L}_{MSE}(G(z_{\mathcal{I}}\ ;\ \theta_G)\ ,\ z_{\mathcal{S}}). \tag{3.6}$$

Algorithm 2 presents the PyTorch-style pseudocode for training the guidance model.

3.2.3 Guided Model F

We incorporate the trained guidance model G into the classification model of the inferior modality $C_{\mathcal{I}}$ so that it is steered to inherit the knowledge captured by the superior classifier $C_{\mathcal{S}}$ but without $C_{\mathcal{I}}$ being exposed to the superior image modality. The predictions of this model, $\hat{y} = F(x_{\mathcal{I}}\ ;\ \theta)$ can be written as follows, with $\theta = \{\theta_{E_{\mathcal{I}}}, \theta_G, \theta_{D_{\mathcal{S}}}\}$:

$$\hat{y} = D_{\mathcal{S}}\left(G\left(E_{\mathcal{I}}(x^i_{\mathcal{I}};\ \theta_{E_{\mathcal{I}}});\ \theta_G\right);\ \theta_{D_{\mathcal{S}}}\right). \tag{3.7}$$

The model F is thus able to make a prediction solely based on the inferior modality while being steered to generate latent representations that mimic those produced by models trained on the superior modality. Here, the superior modality encoder can be viewed as a teacher distilling its knowledge (the learned latent representation) to benefit the encoder of the inferior modality, which would be the student. Figure 3.4 presents an overview of the proposed guided model, F.

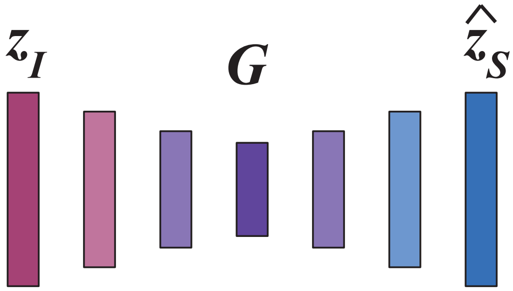


Figure 3.3: The architecture of the guidance model $G(\cdot)$. The guidance model is trained to estimate the latent representation of the superior modality ($\hat{z}_{\mathcal{S}}$) based on the latent representation of the inferior modality ($z_{\mathcal{I}}$).

Algorithm 2 PyTorch-style pseudocode for training the guidance model $G(\cdot)$.

Input : Paired images $\{x^i_{\mathcal{I}}, x^i_{\mathcal{S}}\}\ \forall i \leq N$ and learned encoders $E_{\mathcal{I}}(\cdot)$ and $E_{\mathcal{S}}(\cdot)$ (with parameters $\theta^*_{E_{\mathcal{I}}}$ and $\theta^*_{E_{\mathcal{S}}}$).

Output: Learned guidance model $G(\cdot)$ with parameters θ^*_G.

```
while not converged do
    for $i = 1$ to $N$ do
        ▷ Latent representations
        $z^i_{\mathcal{I}} = E_{\mathcal{I}}(x^i_{\mathcal{I}}\,;\,\theta^*_{E_{\mathcal{I}}})$
        $z^i_{\mathcal{S}} = E_{\mathcal{S}}(x^i_{\mathcal{S}}\,;\,\theta^*_{E_{\mathcal{S}}})$
        ▷ Model predictions
        $\hat{z}^i_{\mathcal{S}} = G(z^i_{\mathcal{I}}\,;\,\theta_G)$
        ▷ Loss computation
        $\mathcal{L} = \mathcal{L}_{MSE}(\hat{z}^i_{\mathcal{S}}, z^i_{\mathcal{S}})$
        ▷ Optimization ($\eta$: learning rate, $\nabla$: gradient)
        $\mathcal{L}$.backward()                    ▷ $\theta^*_G \leftarrow \theta_G - \eta\,\nabla\theta_G$
    end for
end while
```

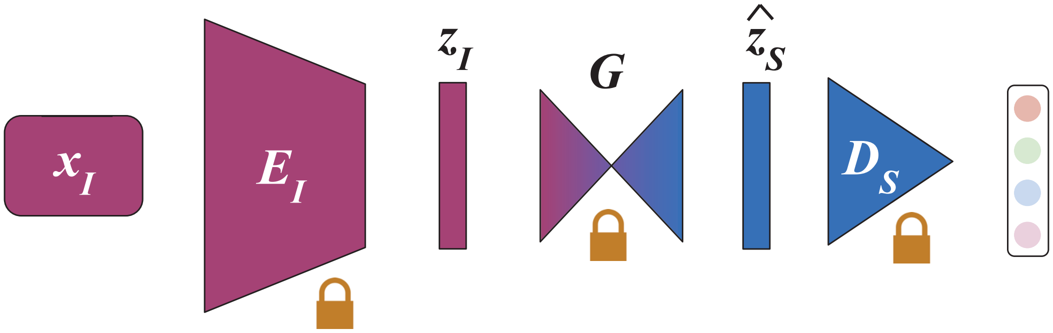


Figure 3.4: G connects the output of the (frozen) inferior modality encoder $E_{\mathcal{I}}$ to the input of the (frozen) superior modality decoder $D_{\mathcal{S}}$. Then G is trained to infer the latent representation of the superior modality from the inferior one.

However, in this model, we use the latent representations of the inferior modality, $z^i_{\mathcal{I}}$ only to produce the latent representations of the superior modality, $\hat{z}^i_{\mathcal{S}}$ that are subsequently used to predict the labels. Essentially, the proposed method predicts the labels based on the latent representations of only the superior modality (produced by the guidance model, G).

In order to leverage the latent representation of the inferior modality, $z^i_{\mathcal{I}}$ during the label prediction, we propose an alternative method for the guided model F as shown in Figure 3.5. The latent representations of the inferior modality, $z^i_{\mathcal{I}}$ are concatenated with the superior latent representations produced by the guidance model, $\hat{z}^i_{\mathcal{S}}$. This ensures the utilization of the knowledge from the inferior modality in making the final prediction.

The concatenated latent representation is used to train a combined decoder $D_c([\hat{z}^i_{\mathcal{S}} \oplus z^i_{\mathcal{I}}]\ ;\ \theta_{D_c})$, with parameters θ_{D_c}, where $\oplus$ denotes the concatenation operator. Thus, our final prediction, $\hat{y} = F(x_{\mathcal{I}}; \theta)$ can be written as follows, with $\theta = \{\theta_{E_{\mathcal{I}}}, \theta_G, \theta_{D_c}\}$:

$$\hat{y} = D_c\left([G\left(E_{\mathcal{I}}(x^i_{\mathcal{I}};\ \theta_{E_{\mathcal{I}}});\ \theta_G\right) \oplus E_{\mathcal{I}}\left(x^i_{\mathcal{I}};\ \theta_{E_{\mathcal{I}}}\right)];\ \theta_{D_c}\right). \tag{3.8}$$

Similar to the classification models discussed in subsection 3.2.1, we use the cross-entropy loss to optimize the classification. We optimize the classification model $F(\cdot)$ as:

$$\theta^*_{D_c} = \underset{\theta_{D_c}}{\arg\min}\, \mathcal{L}_{CE}(F(\mathcal{X}_{\mathcal{I}}\ ;\ \theta)\ ,\ \mathcal{Y}). \tag{3.9}$$

Note that in this case we only optimize the parameters of the decoder (θ_{D_c}) and freeze the parameters of the learned encoder ($\theta_{E_{\mathcal{I}}}$) and the guidance model (θ_G). Algorithm 3 presents the PyTorch-style pseudocode corresponding to this step.

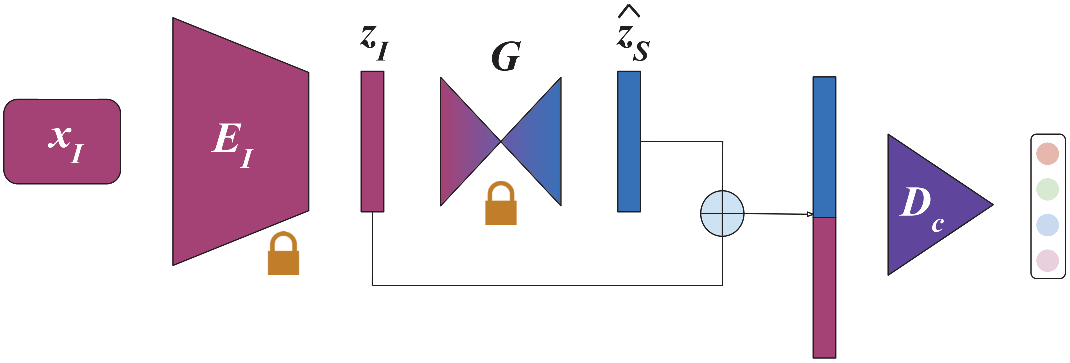


Figure 3.5: The final model, whose input is the inferior modality alone, uses both the inferior and the estimated superior modality representations to make the final prediction via the trained combined decoder D_c.

3.2.4 Inference

The testing images of the inferior modality ($\mathcal{I}$) are passed through the guided model F in order to evaluate the performance of the proposed method. Algorithm 4 presents the PyTorch-style pseudocode corresponding to model inference. Note that we evaluate both variants of the guided model corresponding to Figures 3.4 and 3.5. For convenience, we denote the predictions of the former as $\hat{y}_{G(\mathcal{I})}$ and the latter $\hat{y}_{G(\mathcal{I})+\mathcal{I}}$ in Algorithm 4.

3.3 Loss functions

Cross-entropy loss Given the multi-class classification setup, we make use of the commonly used cross-entropy (CE) loss function in training the classifiers $C_{\mathcal{I}}$ and $C_{\mathcal{S}}$ (Equation 3.1) and also in the final guided model F (Equation 3.8). Denoted by L_{CE}, the CE loss is defined as:

$$\mathcal{L}_{CE}(\hat{\mathcal{Y}}, \mathcal{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} w_c \, y_c^i \log(\hat{y}_c^i), \tag{3.10}$$

where y_c^i and $\hat{y}_c^i$ respectively denote the truth label and the softmax probability prediction of the c-th class of the i-th image, w_c denotes the weight assigned to the c-th class of the label space, and K and N respectively denote the number of class labels and the number of images.

Mean squared error loss Secondly, in order to learn a mapping from the latent representation of the inferior modality ($z_{\mathcal{I}}^i$) to the superior ($z_{\mathcal{S}}^i$), during the training of the guidance model, G, we encourage the estimated latent representations, $\hat{z}_{\mathcal{S}}^i$ to be similar

Algorithm 3 PyTorch-style pseudocode for training the guided model $F(\cdot)$.

Input : Inferior modality images $x^i_{\mathcal{I}}$ and corresponding labels y^i, $\forall i \leq N$. Learned encoder $E_{\mathcal{I}}(\cdot)$ and guidance model $G(\cdot)$ (with parameters $\theta^*_{E_{\mathcal{I}}}$ and θ^*_G).

Output: Learned combined decoder $D_c(\cdot)$ with parameters $\theta^*_{D_c}$.

```
while not converged do
    for $i = 1$ to $N$ do
        ▷ Latent representation
        $z^i_{\mathcal{I}} = E_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta^*_{E_{\mathcal{I}}})$
        ▷ Guided latent representation
        $\hat{z}^i_{\mathcal{S}} = G(z^i_{\mathcal{I}}\ ;\ \theta^*_G)$
        ▷ Concatenation
        $z^i_c = \hat{z}^i_{\mathcal{S}} \oplus z^i_{\mathcal{I}}$
        ▷ Model prediction
        $\hat{y}^i = D_c(z^i_c\ ;\ \theta_{D_c})$
        ▷ Loss computation
        $\mathcal{L} = \mathcal{L}_{CE}(\hat{y}^i,\ y^i)$
        ▷ Optimization ($\eta$: learning rate, $\nabla$: gradient)
        $\mathcal{L}$.backward()                          ▷ $\theta^*_{D_c} \leftarrow \theta_{D_c} - \eta\, \nabla\theta_{D_c}$
    end for
end while
```

Algorithm 4 PyTorch-style pseudocode to evaluate the performance of guided model $F(\cdot)$.

Input : Inferior modality images $x^i_{\mathcal{I}}$ and corresponding labels y^i, $\forall i \leq N_{TEST}$ of the test set. Learned encoder $E_{\mathcal{I}}(\cdot)$, guidance model $G(\cdot)$, and decoders $D_{\mathcal{S}}(\cdot)$ and $D_c(\cdot)$ (with respective parameters $\theta^*_{E_{\mathcal{I}}}$, θ^*_G, $\theta^*_{D_{\mathcal{S}}}$, and $\theta^*_{D_c}$).

Output: Predicted labels $\hat{y}^i_{G(\mathcal{I})}$ and $\hat{y}^i_{G(\mathcal{I})+\mathcal{I}}$, $\forall i \leq N_{TEST}$.

```
for $i = 1$ to $N_{TEST}$ do
    ▷ Latent representation
    $z^i_{\mathcal{I}} = E_{\mathcal{I}}(x^i_{\mathcal{I}}\ ;\ \theta^*_{E_{\mathcal{I}}})$
    ▷ Guided latent representation
    $\hat{z}^i_{\mathcal{S}} = G(z^i_{\mathcal{I}}\ ;\ \theta^*_G)$
    ▷ Model prediction
    $\hat{y}^i_{G(\mathcal{I})} = D_{\mathcal{S}}(\hat{z}^i_{\mathcal{S}}\ ;\ \theta^*_{D_{\mathcal{S}}})$                 ▷ Guided model of Figure 3.4
    ▷ Concatenation
    $z^i_c = \hat{z}^i_{\mathcal{S}} \oplus z^i_{\mathcal{I}}$
    ▷ Model prediction
    $\hat{y}^i_{G(\mathcal{I})+\mathcal{I}} = D_c(z^i_c\ ;\ \theta^*_{D_c})$                 ▷ Guided model of Figure 3.5
end for
```

to the latent representations of the superior modality, $z^i_{\mathcal{S}}$. We use the mean squared error (MSE) loss function for this purpose. Denoted by L_{MSE}, the MSE loss is defined as:

$$\mathcal{L}_{MSE}(\hat{z}, z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}(z^i_j - \hat{z}^i_j)^2, \tag{3.11}$$

where z^i and $\hat{z}^i$ respectively denote the expected and the predicted latent representations of the i-th image, N denotes the number of images, and M denotes the dimensionality of the latent representations.

Chapter 4

Experimental setup

4.1 Datasets

We evaluate the proposed method on two multimodal image datasets that correspond to the diseases of two different body organs and vastly different image dimensions. Our first dataset is RadPath 2020, a brain tumor classification dataset consisting of multi-sequence MRI and histopathology WSI. The second dataset is Derm7pt, a multi-task skin lesion classification dataset with clinical and dermoscopic images of skin lesions. The following subsections provide a detailed description of the two datasets.

4.1.1 RadPath 2020

RadPath 2020 [14], or simply RadPath, is a publicly available dataset that was released as part of the Computational Precision Medicine Radiology-Pathology (CPM RadPath) Challenge of 2020. The challenge was part of the Brain-Lesion Workshop [82], a satellite event of the 2020 MICCAI conference.

RadPath is a brain tumor classification dataset that consists of 221 pairs of multi-sequence MRI and histopathology WSI, along with the diagnosis labels of the corresponding patients. The diagnosis labels include the three major types of glioma – glioblastoma ($n = 133$), oligodendroglioma ($n = 34$), and astrocytoma ($n = 54$). As shown in Table 4.1, we divide the dataset into training, validation, and testing sets with respectively 165, 28, and 28 image pairs. Due to the small size of the dataset, we make five such splits and present the average performance across all the splits for robust results.

Figure 4.1 shows an example of an image pair from the RadPath dataset. The multi-sequence MRI consists of four sequences – native (T1), T2-weighted (T2), post-contrast T1-weighted (T1Gd), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). Each MRI sequence is a volumetric image with the size $240 \times 240 \times 155$. All MRI scans in the dataset are co-registered to the same anatomical template, interpolated to the same resolution,

Table 4.1: Details of the RadPath dataset. The table shows the distribution of data across the classes and the training, validation, and testing sets.

Class	# of image pairs			
	train	val	test	total
Glioblastoma	103	15	15	133
Oligodendroglioma	24	5	5	34
Astrocytoma	38	8	8	54
Total	165	28	28	221

and skull-stripped by the providers of the dataset. The histopathology data consists of one WSI for each patient. The WSIs are digitized Hematoxylin and Eosin (H&E) stained tissue specimens scanned at 20× or 40× magnifications. These images are of extremely high resolution with dimensions of some WSIs as high as $3 \times 100,000 \times 100,000$. The relatively smaller WSIs in the dataset are as large as $3 \times 40,000 \times 40,000$. As discussed previously in section 1.2, the WSIs, due to the level of informativeness and invasive acquisition procedures, form the superior modality, whereas the MRIs form the inferior modality due to the relative ease of acquisition and lower informativeness regarding the tumor.

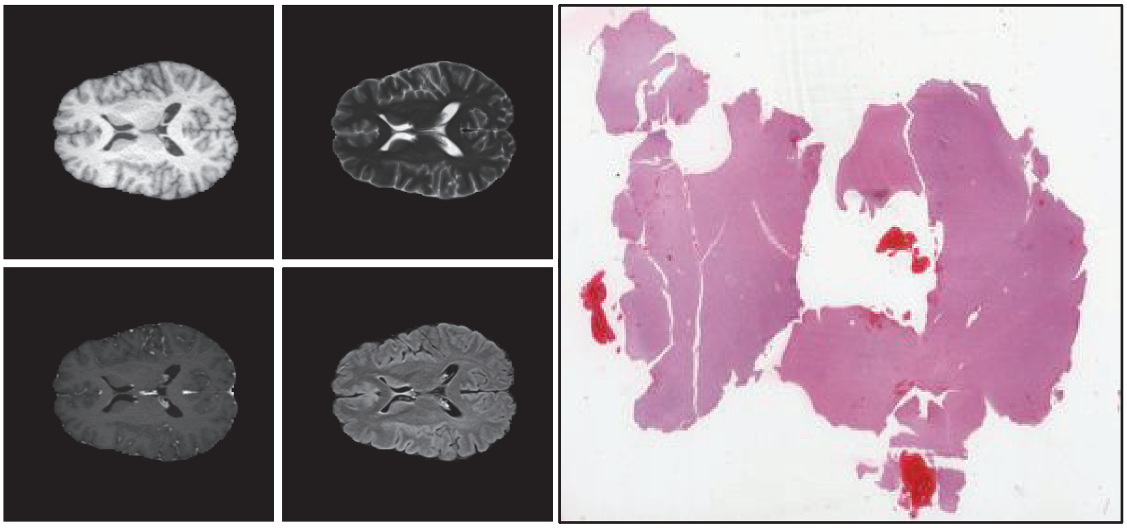


Figure 4.1: Example of an image pair from the RadPath dataset belonging to the Oligodendroglioma class. The image pair consists of multi-sequence MRI (*left*) of the brain and high-resolution WSI of the tumor tissue (*right*) from the same patient. The multi-sequence MRI contains four sequences: T1 (*top-left*), T2 (*top-right*), T1Gd (*bottom-left*), and T2-FLAIR (*bottom-right*).

4.1.2 Derm7pt

Derm7pt [15] is a publicly available dataset consisting of paired skin lesion images from the clinical and dermoscopic modalities. The dataset includes 1011 pairs of images, where both the images of the image pair correspond to the same skin lesion acquired from the same patient. Each image-pair consists of a diagnosis label along with seven 7-point criteria labels, allowing for multiple ($n = 8$) classification tasks using this dataset. Table 4.2 shows the labels of diagnosis and the 7-point criteria and the distribution of data across the different labels. In our experiments, we adopt the pre-defined training, validation, and testing sets provided with the dataset and repeat our training procedure three times to present results that are robust to different random weight initializations. We observed that 8 image pairs in the dataset are missing the clinical image and therefore we exclude them from our experiments.

The 7-point criteria [83] is an analytical, rule-based method used to simplify the detection of melanoma. The 7-point criteria comprise seven prominent features of skin lesions that are frequently associated with melanoma, namely, pigment network (PN), blue whitish veil (BWV), vascular structures (VS), pigmentation (PIG), streaks (STR), dots and globules (DaG), and regression structures (RS). Each criterion is assigned a score based on its association with melanoma and if the cumulative score across all seven criteria exceeds a given threshold, the lesion is diagnosed as melanoma. Table 4.2 shows the 7-point scores assigned to the 7-point criteria. PN, BWV, and VS – considered major criteria in the detection of melanoma – are assigned a score of 2, while the remaining four (PIG, STR, DaG, and RS) are considered minor and are assigned a score of 1.

The clinical and dermoscopic images in this dataset are 2D images of size $3 \times 512 \times 512$. The dermoscopic images are acquired through a dermoscope and reveal detailed sub-surface structures of the skin lesion that are not visible to the unaided eye. As discussed previously in section 1.3, the dermoscopic images form the superior modality, whereas the clinical images, acquired using inexpensive and ubiquitous cameras, form the inferior modality. Figure 4.2 shows three image pairs from the Derm7pt dataset. The clinical images are placed on the top row and the corresponding dermoscopic images are on the bottom row.

4.2 Data processing

4.2.1 RadPath

4.2.1.1 WSI processing

We make use of CLAM [84] – CLustering-constrained Attention-based MIL – a recently proposed DL-based weakly-supervised learning framework for the classification of WSIs.

Table 4.2: Details of the Derm7pt dataset. The table shows the different classification tasks (diagnosis and the 7-point criteria), the data distribution across labels of different tasks, and the 7-point scores corresponding to labels of 7-point criteria used to infer melanoma.

Class	# of images	7-point score
	DIAGNOSIS (DIAG)	
Basal Cell Carcinoma	42	-
Nevus	575	-
Melanoma	252	-
Miscellaneous	97	-
Seborrheic Keratosis	45	-
	SEVEN POINT CRITERIA	
	1. Pigment Network (PN)	
Absent	400	0
Typical	381	0
Atypical	230	2
	2. Blue Whitish Veil (BWV)	
Absent	816	0
Present	195	2
	3. Vascular Structures (VS)	
Absent	823	0
Regular	117	0
Irregular	71	2
	4. Pigmentation (PIG)	
Absent	588	0
Regular	118	0
Irregular	305	1
	5. Streaks (STR)	
Absent	653	0
Regular	107	0
Irregular	251	1
	6. Dots and Globules (DaG)	
Absent	229	0
Regular	334	0
Irregular	448	1
	7. Regression Structures (RS)	
Absent	758	0
Present	253	1

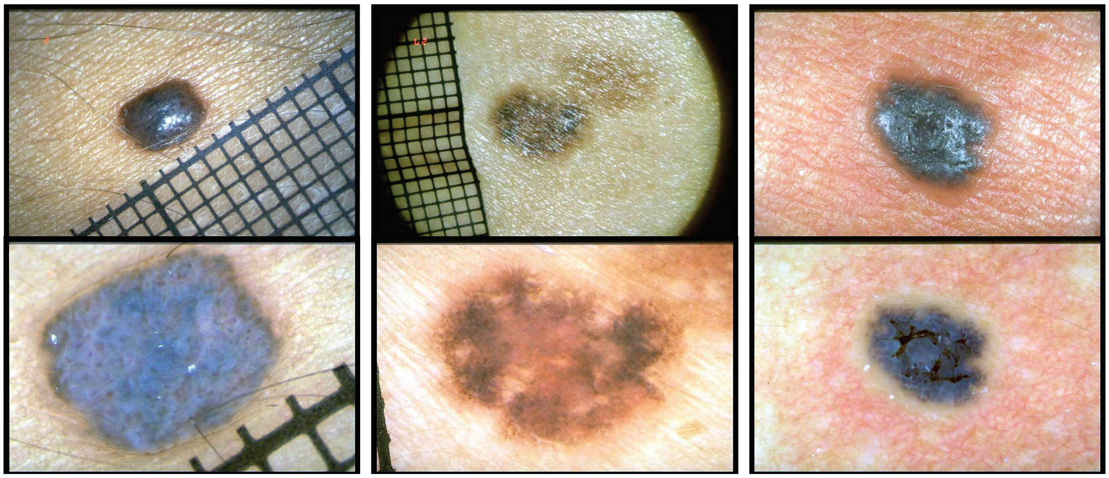


Figure 4.2: Examples of three image-pairs from Derm7pt dataset belonging to the diagnosis classes of basal cell carcinoma (*left*), nevus (*middle*), and melanoma (*right*) respectively. The clinical images are placed on the top row and the corresponding dermoscopic images are on the bottom row. Notice the sub-cellular structures visible in the dermoscopic images but not in the clinical counterparts.

Specifically, this model serves as the modality-specific classifier for the superior modality, $C_{\mathcal{S}}(\cdot)$ as per Equation 3.1 and comprises encoder $E_{\mathcal{S}}(\cdot)$ and decoder $D_{\mathcal{S}}(\cdot)$ as shown in Figure 3.2. CLAM uses an attention-based learning strategy to automatically identify regions of tumorous tissues while using only the slide-level labels and without requiring any manual annotations. This method has been shown to outperform other standard weakly-supervised WSI classification methods, especially in cases where the size of training data is small. In our work, we use the same preprocessing pipeline and the training strategy proposed in CLAM to train the WSI classifier for the RadPath dataset.

The preprocessing pipeline proposed in CLAM involves three steps:

1. Segmentation of tissue regions of the WSI,
2. Patching the segmented areas into images of smaller dimensions, and
3. Feature extraction from the image patches.

Figure 4.3 shows an overview of the segmentation and the subsequent patching of the WSI.

In the first step, the WSIs are converted from RGB to HSV color space and the binary segmentation masks are computed based on the thresholding of the saturation channel. The obtained segmentation masks are filtered based on a tissue-area threshold and the holes are filtered based on a hole-area threshold. Due to the variability in the sizes of tissues and

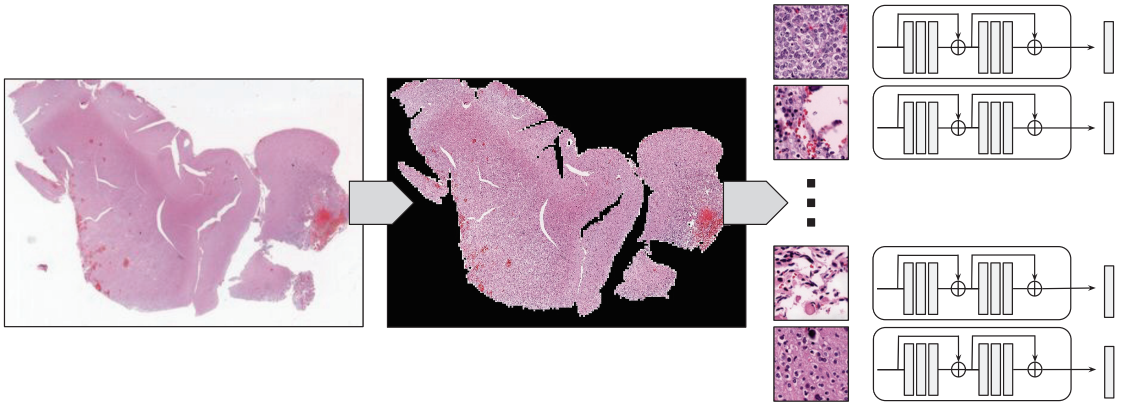


Figure 4.3: WSI preprocessing. A raw WSI is first segmented with tissue regions on the foreground, followed by patching of the segmented tissue region into patches of 256 × 256 dimension. The extracted tissue patches are passed through a ResNet50 model pre-trained on ImageNet to extract a 1024-dimension representation for each patch.

holes across different WSIs, these thresholds are to be manually adjusted for different WSIs.

The segmented tissue regions of the WSI are then cropped exhaustively at 20× magnification with patch sizes of 256 × 256. As our dataset contains some WSIs at 40× magnification, we crop patches of size 512 × 512 for such WSIs and downscale these patches to 256 × 256 in order to have patches at 20× magnification. Depending on the sizes of WSIs and the tissue content, the number of patches per WSI can vary significantly across different WSIs.

Finally, the extracted patches of all WSIs are passed through a ResNet50 model pre-trained on ImageNet to convert each 256×256 patch into a 1024-dimension representation. These low-dimension image representations are used to train the CLAM model, which allows for large WSIs with hundreds of thousands of patches to fit into the GPU memory.

A WSI, W^i can now be represented as a bag of low-dimension patch representations, z_p of all the patches belonging to the corresponding WSI; $W^i = \{z^i_{p_1}, z^i_{p_2}, \ldots, z^i_{p_n}\}$, where $z^i_{p_j}$ represents the low-dimension patch representation of patch p_j of WSI W^i and n is the number of patches in W^i (n is different for different WSIs). With these WSI bags as input, CLAM uses an MIL framework to first obtain a WSI-level representation, z^i_W from the patch-level representations in the bag, followed by using the obtained WSI representation to predict the diagnosis label assigned to the WSI in an end-to-end fashion.

The patch representations are first passed through a learnable FC layer to reduce the dimension from 1024 to 512. The WSI-level representation of W^i, z^i_W, is then computed as the linear combination of all the patch-level representations as, $z_W^i = \sum_{j=1}^{n} a_j^i \, z_{p_j}^i$, where $a_j^i \in \mathbb{R}[0,1]$ is the learned attention score of the patch corresponding to $z_{p_j}^i$. Note that after the first FC layer, $z_{p_j}^i \in \mathbb{R}^{512}$, so $z_W^i \in \mathbb{R}^{512}$. The attention score, a_j^i – which is the relative importance of the patch p_j with respect to all other patches in W^i – is computed as, $a_j^i = softmax_j\{A(z_{p_j}^i)\}$, where $A(\cdot)$ is the attention network composed of three FC layers which regress to a scalar value that denotes the importance of the patch. Finally, the WSI-level representation, z_W^i is used to predict the diagnosis label corresponding to the WSI via FC layers. For a more detailed methodology of CLAM, we direct the reader to the respective publication by Lu *et al.* [84].

4.2.1.2 MRI processing

As mentioned previously in subsection 4.1.1, the radiology data of the RadPath challenge consists of preprocessed 3D MRIs. The MRIs are co-registered to the same anatomical template, interpolated to the same resolution, and skull-stripped by the providers to the dataset. Apart from normalizing the intensities of the individual MRI sequences (on the fly), we do not use any other preprocessing techniques in our experiments. In order to preserve the original high resolution, we use the MRIs at sizes $240 \times 240 \times 155$ as inputs to our network, without any downscaling.

Similar to the winners of the RadPath 2019 challenge [49], we employ a 3D DenseNet to process the volumetric MRIs. 3D DenseNet [85] has been the go-to choice of the backbone network architecture for a majority of MRI classification models in the RadPath challenge [86, 87] – the winners of the RadPath 2020 challenge [48] also used the same backbone network for MRI classification. This model serves as the modality-specific classifier for the inferior modality, $C_{\mathcal{I}}(\cdot)$ as per Equation 3.1 and comprises encoder $E_{\mathcal{I}}(\cdot)$ and decoder $D_{\mathcal{I}}(\cdot)$ as shown in Figure 3.2.

Our implementation of the MRI classifier closely follows that of the winners of the RadPath 2019 challenge [49]. We train our network from scratch and make use of excessive image augmentations to compensate for the small size of the training dataset. Our augmentations include random flipping and rotation of the MRI sequences along all three axes, random scaling and cropping of the MRI sequences, and random scaling of the intensities of the MRI sequences.

In our experiments, we train four different models, one for each of the MRI sequences – T1, T2, T1Gd, and T2-FLAIR – and *guide* each of these models using the superior modality (WSI). We also train a model with all the MRI sequences where we use the pre-trained encoders of each of the four sequence-specific models to first extract the latent representation of each sequence followed by training a decoder (made of FC layers) on the concatenated latent representation.

Finally, for our baseline multimodal model, which uses MRI and WSI for classification, we use the pre-trained encoders of MRI and WSI classification models and concatenate the two latent representations. Based on this concatenated representation, we train a decoder to classify the brain tumors in the MRI-WSI image pairs.

4.2.2 Derm7pt

Kawahara *et al.* [15] proposed a multimodal skin lesion classification model that uses images from both the clinical and dermoscopic modalities along with the patient meta-data to classify skin lesions. Their multimodal model uses two Inception V3 [88] backbone networks to extract latent representations from the clinical and dermoscopic modalities. These latent representations, along with the patient meta-data, are combined to perform a multimodal classification. Additionally, their multimodal model is trained to handle all possible combinations of input modalities and as a result, it is capable of making predictions with missing data at the inference time.

In our experiments, we use this pre-trained multimodal model as a baseline classifier for both the clinical and dermoscopic modalities. While evaluating our clinical classifier, we pass only the clinical image as input to the multimodal model while passing zeros at the dermoscopic and the meta-data inputs. Similarly, while evaluating our dermoscopic model, we pass the dermoscopic image as the input while passing zeros at the clinical and meta-data inputs of the multimodal model. In both cases, similar to Kawahara *et al.* [15], we preprocess the image inputs as per the Inception V3 input requirements – using the in-built Tensorflow [89] function `tensorflow.keras.applications.inception_v3.preprocess_input`.

The pre-trained clinical model forms the modality-specific classifier for the inferior modality ($C_{\mathcal{I}}(\cdot)$, comprising of $E_{\mathcal{I}}(\cdot)$ and $D_{\mathcal{I}}(\cdot)$) and the pre-trained dermoscopic model forms the modality-specific classifier for the superior modality ($C_{\mathcal{S}}(\cdot)$, comprising of $E_{\mathcal{S}}(\cdot)$ and $D_{\mathcal{S}}(\cdot)$) as in Equation 3.1 and Figure 3.2.

For our baseline multimodal model – which uses both clinical and dermoscopic images to classify skin lesions – we follow a similar technique as above. We use the aforementioned pre-trained multimodal model provided by Kawahara *et al.* [15] and pass both the inputs to the model while passing zeros at the meta-data input.

4.3 Implementation details

For our experiments on the RadPath dataset, we use the PyTorch [90] library and for our experiments on the Derm7pt dataset, we use the Tensorflow [89] library, to build on top of the Keras-based pre-trained models provided by Kawahara *et al.* [15]. Additionally, we use the OpenSlide [91] library for preprocessing the WSIs, specifically for the segmentation and patching of the WSIs, discussed in section 4.2.1.1. For loading the MRI volumes and the associated 3D data augmentations, we use the MONAI [92] library. For evaluating the performance of the classifications, we use the metrics implemented in the Scikit-Learn library [93].

In both sets of experiments, our guidance model uses an autoencoder-like architecture as described in section 3.2.2. We use a bottleneck layer with 256 neurons for our guidance model in the RadPath dataset and a bottleneck layer with 512 neurons for the same in the Derm7pt dataset. Specifically, in the RadPath dataset, we map from a 1024-dimension latent space of the inferior modality (MRI) to the 512-dimension latent space of the superior modality (WSI) with FC layers as $1024 \rightarrow 512 \rightarrow 256 \rightarrow 256 \rightarrow 512$. Similarly, in the Derm7pt dataset, we map from a 2048-dimension latent space of clinical modality to a 2048-dimension latent space of dermoscopic modality as $2048 \rightarrow 1024 \rightarrow 512 \rightarrow 1024 \rightarrow 2048$. In both cases, we use the ReLU non-linearity and use Dropout layers with a neuron drop rate of 25% to prevent overfitting.

With respect to the computing hardware, we make use of the multi-GPU cluster provided by Compute Canada [94] for the resource-intensive jobs in our experiments. Specifically, all our experiments on WSIs, starting from the preprocessing to the training of the classification model, and most of our experiments on MRIs, including training the classification models utilized the resources provided by Compute Canada. However, our experiments on the Derm7pt dataset relied on the local 11 GB NVIDIA GeForce GTX 1080 Ti GPU.

Table 4.3 provides the values of the optimal hyperparameters used in different experiments across the two datasets. The table includes the common hyperparameters such as the batch size, number of epochs, optimizer parameters, early stopping parameters, loss weights, and weighted random sampling. For further details on the implementation, we direct the reader to our GitHub repository at `https://github.com/mayurmallya/DeepGuide`.

Table 4.3: Optimal values of the hyperparameters used in our experiments across RadPath and Derm7pt datasets. The tasks in RadPath experiments include T1, T2, T1Gd, T2-FLAIR, and ALL. The tasks in Derm7pt experiments include PN, BWV, VS, PIG, STR, DaG, RS, and DIAG. The abbreviations used are as follows: BS: batch size, optim: optimizer, LR: learning rate, WRS: weighted random sampler, '$*$': all tasks.

Experiment	task	GPU	BS	epochs	patience	network	optim	LR	loss	loss weights	WRS
$P + R$	-	11G	50	500	200	FC	SGD	1×10^{-3}	CE	[1.0, 1.7, 1.6]	✓
P	-	4×12G	1	200	20	CLAM	Adam	1×10^{-4}	CE	-	✓
R	$*$	32G	10	600	50	DenseNet121	Adam	2×10^{-4}	CE	-	✓
$G(R)$	$*$	11G	50	150	-	FC	SGD	5×10^{-1}	MSE	-	-
$G(R) + R$	$*$	11G	50	500	200	FC	SGD	1×10^{-3}	CE	[1.0, 1.7, 1.6]	✓
$D + C$	$*$	Pre-trained model									
D	$*$	Pre-trained model									
C	$*$	Pre-trained model									
$G(C)$	$*$	11G	100	500	-	FC	SGD	5×10^{-1}	MSE	-	-
$G(C) + C$	PN	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.9, 0.9, 1.8]	-
	BWV	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.6, 12.0]	-
	VS	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.01, 20.0, 50.0]	-
	PIG	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.2, 45.0, 1.1]	-
	STR	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.06, 50.0, 4.0]	-
	DaG	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.3, 0.9, 0.9]	-
	RS	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[0.25, 6.0]	-
	DIAG	11G	100	500	200	FC	SGD	1×10^{-4}	CE	[15.0, 0.07, 1.4, 4.0, 15.0]	-

4.4 Evaluation metrics

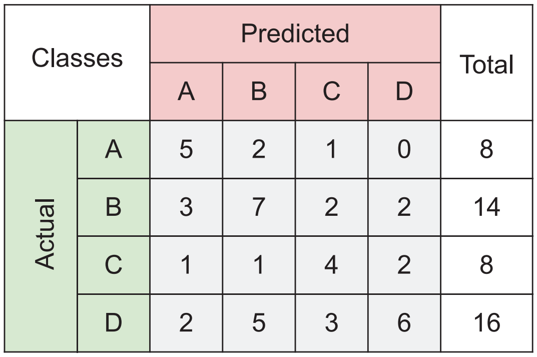
Our choice of evaluation metrics is inspired by the RadPath challenge [14], which uses Balanced Accuracy, Micro-averaged F1, and Cohen's Kappa scores to measure the classification performance. In our experiments, we use balanced accuracy along with micro-averaged F1. However, we do not use the kappa score due to the undesired behaviors this metric exhibits in cases of unbalanced classification [95] and the resulting lack of interpretability. For the binary task of melanoma inference, we use the AUROC score to measure the classification performance.

Balanced accuracy

As our experiments mostly involve unbalanced classification tasks, we use the balanced accuracy score as the primary metric to evaluate the classification performance. Balanced accuracy is calculated as the macro-average of class-wise accuracies and gives equal importance to all the classes irrespective of the class sizes. If we denote the balanced accuracy score as BA, equation 4.1 shows the corresponding formulation:

$$BA = \frac{\sum\limits_{i=1}^{N} \frac{TP_i}{Total_i}}{N}. \tag{4.1}$$

Here N denotes the number of classes, TP denotes the number of true positives, and $Total$ denotes the total number of elements in the particular class. The TP of a particular class is the value on the confusion matrix at the intersection of the row and column corresponding to that class. In the example confusion matrix shown in figure 4.4, the balanced accuracy can be calculated as:



Classes		Predicted				Total
		A	B	C	D	
Actual	A	5	2	1	0	8
	B	3	7	2	2	14
	C	1	1	4	2	8
	D	2	5	3	6	16

Figure 4.4: Example of a confusion matrix.

$$BA = \frac{\frac{5}{8} + \frac{7}{14} + \frac{4}{8} + \frac{6}{16}}{4} = \frac{0.625 + 0.5 + 0.5 + 0.375}{4} = 0.5.$$

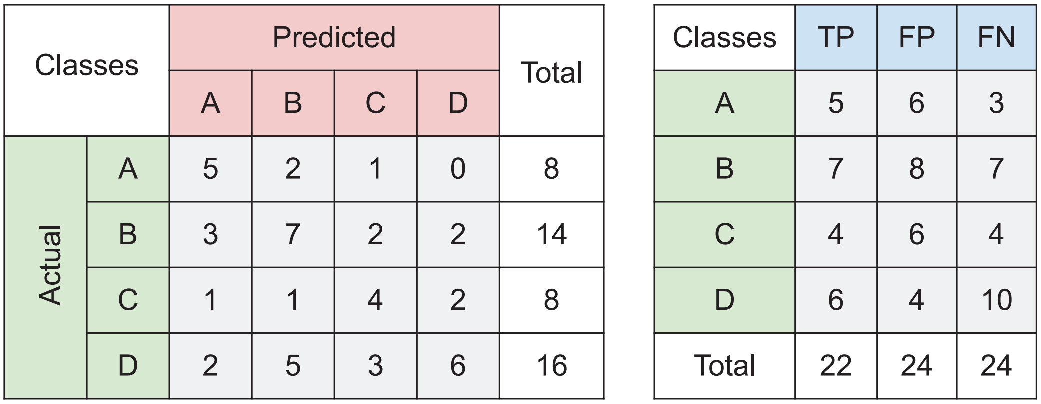
Although balanced accuracy is a commonly used metric for evaluating classifications on unbalanced datasets, a consequence of the equal weighting of class accuracies is that the minority classes have more influence over the final score. For example, a misclassification in class D (majority class) would only set back the score by $\frac{1}{4}(\frac{1}{16})$, whereas a misclassification in class A (minority class) would set back the score by $\frac{1}{4}(\frac{1}{8})$ – which is twice as that of the former. In a situation with high class imbalance, this leads to a balanced accuracy score that is less sensitive to predictions of the majority class. In order to circumvent this problem, we use the micro-averaged F1 score in conjunction with the balanced accuracy score.

Micro-averaged F1

In addition to balanced accuracy, we also use the micro-averaged F1 score to provide a holistic view of the classification performance. Unlike balanced accuracy, which gives equal importance to all classes, micro F1 represents the overall correctness of the classifier, irrespective of the class performance, and gives equal importance to the individual elements. By using this metric alongside balanced accuracy, we ensure that our metrics are sensitive to the majority classes too. Micro-averaged F1 score is computed as the harmonic mean of micro-averaged precision and recall scores. Denoting micro F1 as simply F1, equation 4.2 shows the corresponding formulation:

$$F1 = 2\left(\frac{P \cdot R}{P + R}\right), \tag{4.2}$$

where P is the micro-averaged precision score given by, $P = \frac{TP}{TP+FP}$ and R is the micro-averaged recall score given by, $R = \frac{TP}{TP+FN}$. Here TP, FP, and FN are true positives, false positives, and false negatives respectively.



Classes		Predicted				Total
		A	B	C	D	
Actual	A	5	2	1	0	8
	B	3	7	2	2	14
	C	1	1	4	2	8
	D	2	5	3	6	16

Classes	TP	FP	FN
A	5	6	3
B	7	8	7
C	4	6	4
D	6	4	10
Total	22	24	24

Figure 4.5: Example of a confusion matrix (*left*) and the corresponding table with class-wise distribution of true positives (TP), false positives (FP), and false negatives (FN) (*right*).

Figure 4.5 shows an example of a confusion matrix along with a table that computes the class-wise TP, FP, and FN of the corresponding confusion matrix. To give an example, TP of class B is the value at the intersection of column B and row B, FP of class B is the sum of the rest of the values in column B, and FN of class B is the sum of rest of the values in row B. The F1 of the below confusion matrix can be computed as follows:

$$F1 = 2\left(\frac{\frac{22}{22+24}\cdot\frac{22}{22+24}}{\frac{22}{22+24}+\frac{22}{22+24}}\right) = \frac{22}{22+24} = \frac{22}{46} = 0.478.$$

In the case of a multi-class classification where each element has exactly one label (as in ours), FP is equal to FN (Figure 4.5), which is equal to the sum of non-diagonal elements of the confusion matrix. As a result, the F1 score gets the same value as that of the accuracy of the classifier [96].

AUROC

For the binary task of melanoma inference, we use the AUROC score. AUROC stands for the Area Under the Receiver Operating Characteristics curve and is a metric that summarizes the performance of a probabilistic binary classifier across different thresholds. Figure 4.6 shows an example ROC curve (denoted in blue) plotted across different TPR and FPR coordinates obtained from a binary classifier tested using different values of threshold. TPR is the true positive rate of the classifier and is given by, $TPR = \frac{TP}{TP+FN}$ and FPR is the false positive rate given by, $FPR = \frac{FP}{FP+TN}$.

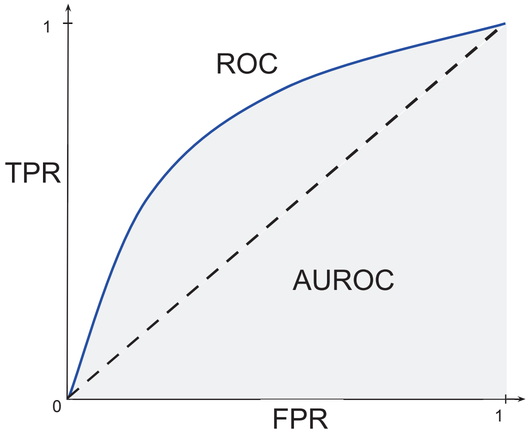


Figure 4.6: Example of an ROC curve. The area of the shaded region under the ROC curve corresponds to the AUROC score. The dashed line corresponds to the ROC curve of a random classifier with an AUROC score of 0.5.

Chapter 5

Results and Discussion

In this chapter, we present the results of our experiments on RadPath (section 5.1) and Derm7pt (section 5.2) datasets and discuss the findings of our experiments (section 5.3). In our experiments, we compare the performance of the proposed method to multiple different baselines. Specifically, we start with the multimodal model – a model that uses multimodal data of the superior and inferior modalities for both training and inference. Figure 5.1 shows the schematic of the multimodal model used in our experiments. Next, we have the modality-specific models for the superior and inferior modalities. As shown in Figure 3.2, the superior modality classifier uses the superior modality images for training and inference, and similarly, the inferior modality model uses the inferior modality data for both training and inference. Finally, we compare the performance of the aforementioned models to the proposed method that uses both superior and inferior modalities for training but only the inferior modality during inference (Figure 3.4 and Figure 3.5).

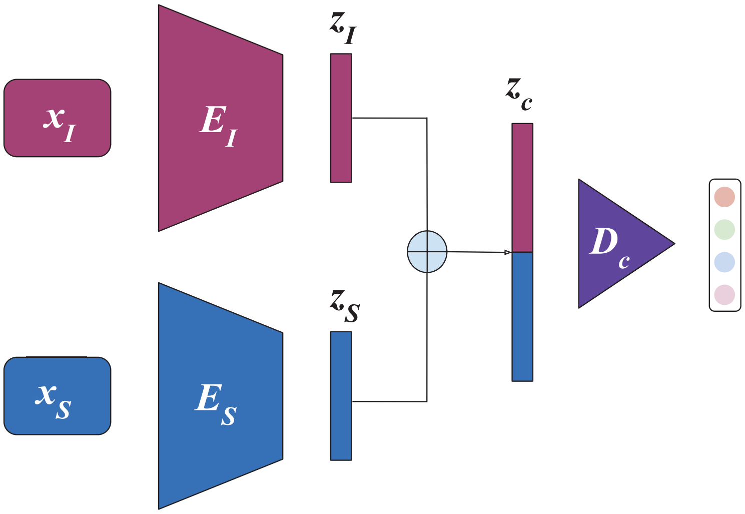


Figure 5.1: Multimodal model that uses the superior and inferior modality data for both training and inference.

5.1 RadPath

Table 5.1 shows the results of our experiments on the RadPath dataset. Note that radiology (denoted by R) forms the inferior modality and histopathology (denoted by P) forms the superior modality in this case. We denote the different models based on the modalities used for the task. Accordingly, the multimodal model can be denoted by $P + R$, while the modality-specific models of histopathology and radiology can be respectively denoted by P and R. Further, we denote the guided models using $G(R)$ and $G(R) + R$, where G, as defined in section 3.2.2, refers to the guidance model. The improvement in the performance of the guided model, $G(R) + R$ over the baseline radiology model, R is denoted by Δ_R. If we denote the performances of the guided model and the radiology model by $\mathcal{X}_{G(R)+R}$ and $\mathcal{X}_R$ respectively, then Δ_R is computed as follows:

$$\Delta_R = \frac{\mathcal{X}_{G(R)+R} - \mathcal{X}_R}{\mathcal{X}_R} \times 100. \tag{5.1}$$

Similarly, Δ_P shows the improvement in the performance of the guided model, $G(R)+R$ over the baseline histopathology model, P, and can be computed as:

$$\Delta_P = \frac{\mathcal{X}_{G(R)+R} - \mathcal{X}_P}{\mathcal{X}_P} \times 100. \tag{5.2}$$

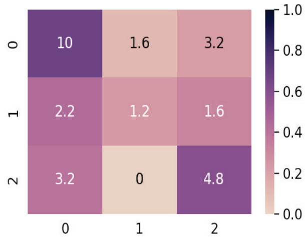
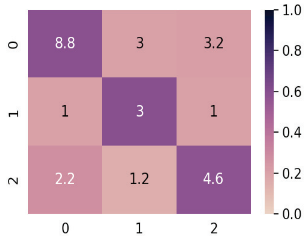
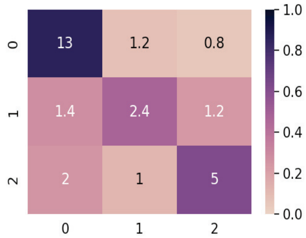
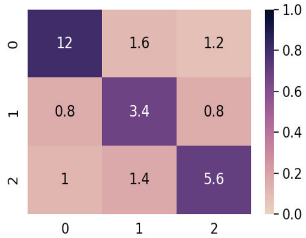
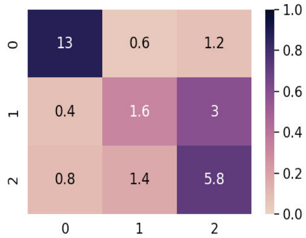
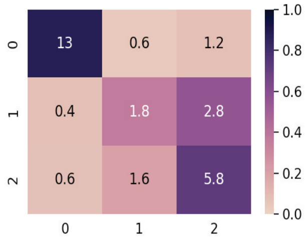
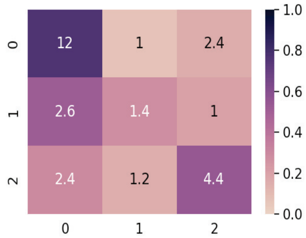
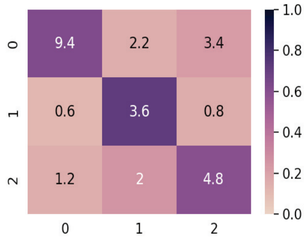
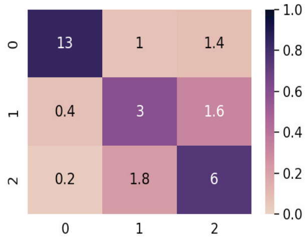
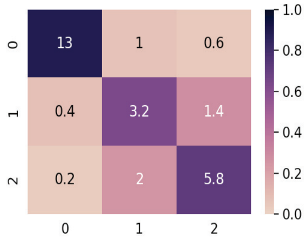
For the baseline modality-specific classifiers of radiology (R) and histopathology (P), we implement the state-of-the-arts respectively using the 3D DenseNet [49] and CLAM models [84], discussed in section 4.2. The multimodal model (P+R) uses the pre-trained encoders of both modalities and only trains a decoder to map the combined latent representation to the corresponding labels.

As previously mentioned in section 4.1.1, the radiology data consists of multi-sequence MRI with the sequences T1, T2, T1Gd, and T2-FLAIR. In our experiments, we use the histopathology images to initially *guide* the individual sequences before finally *guiding* the multi-sequence MRI. The check marks in Table 5.1 denote the sequences of MRI used in the corresponding experiments. Further, Table 5.2 shows the confusion matrices denoting the improvements of the guided model ($G(R) + R$) over the radiology model (R).

Table 5.1: Results on the RadPath dataset. Radiology MRI sequences (R) are guided using pathology (P). R: Inferior, P: Superior. The metrics are balanced accuracy (BA) and micro-averaged F1 scores.

Method	Sequences of R				BA↑	F1↑
	T1	T2	T1Gd	T2-FLAIR		
$P + R$	✓	✓	✓	✓	0.7777 ±0.0697	0.8213 ±0.0319
P [84]	-	-	-	-	0.7293 ±0.0276	0.7928 ±0.0142
R [49]	✓	-	-	-	0.5065 ±0.0348	0.5785 ±0.0692
$G(R)$	✓	-	-	-	0.4432 ±0.0668	0.5784 ±0.0614
$R + G(R)$	✓	-	-	-	0.5871 ±0.0295	0.5856 ±0.0428
Δ_R (%)					+16.0	+1.2
R [49]	-	✓	-	-	0.6571 ±0.0710	0.7285 ±0.0428
$G(R)$	-	✓	-	-	0.5599 ±0.1268	0.6928 ±0.0940
$G(R) + R$	-	✓	-	-	0.7310 ±0.0859	0.7571 ±0.0728
Δ_R (%)					+11.2	+3.9
R [49]	-	-	✓	-	0.6416 ±0.0408	0.7356 ±0.0175
$G(R)$	-	-	✓	-	0.5649 ±0.0728	0.6928 ±0.0622
$G(R) + R$	-	-	✓	-	0.6549 ±0.0431	0.7428 ±0.0143
Δ_R (%)					+2.0	+0.9
R [49]	-	-	-	✓	0.5344 ±0.0488	0.6213 ±0.0484
$G(R)$	-	-	-	✓	0.4199 ±0.0522	0.5713 ±0.0319
$G(R) + R$	-	-	-	✓	0.6488 ±0.0327	0.6356 ±0.0524
Δ_R (%)					+21.3	+2.2
R [49]	✓	✓	✓	✓	0.7299 ±0.0655	0.7716 ±0.0365
$G(R)$	✓	✓	✓	✓	0.6505 ±0.0683	0.7642 ±0.0364
$G(R) + R$	✓	✓	✓	✓	0.7527 ±0.0806	0.7999 ±0.0285
Δ_R (%)					+3.1	+3.6
Δ_P (%)					+3.2	+0.89

Table 5.2: Confusion matrices of the inferior modality model ($\mathcal{I}$) and the proposed guided model ($G(\mathcal{I}) + \mathcal{I}$) for the RadPath dataset. The arrows (↑, ↓) indicate the direction of improvement in the balanced accuracy scores.

<table>
<tr><th>Sequences of R</th><th>w/o guidance</th><th>w/ guidance</th></tr>
<tr><td>T1(↑)</td><td>
</td><td>
</td></tr>
<tr><td>T2(↑)</td><td>
</td><td>
</td></tr>
<tr><td>T1Gd(↑)</td><td>
</td><td>
</td></tr>
<tr><td>T2-FLAIR(↑)</td><td>
</td><td>
</td></tr>
<tr><td>All(↑)</td><td>
</td><td>
</td></tr>
</table>

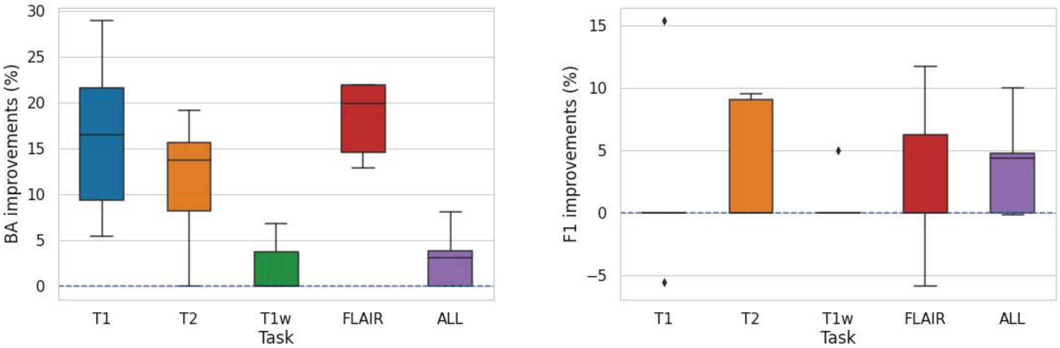


Figure 5.2: Improvements in the balanced accuracy (*left*) and F1 (*right*) scores of the guided model ($G(\mathcal{I})+\mathcal{I}$) over the baseline model (I) for the individual sequences of MRI (T1, T2, T1Gd, and T2-FLAIR) and multi-sequence MRI denoted as ALL.

5.1.1 Statistical significance

We use the Wilcoxon signed-rank test to evaluate the statistical significance of the performance of the proposed method ($G(\mathcal{I})+\mathcal{I}$) over the baseline inferior modality model ($\mathcal{I}$) of the RadPath dataset. Wilcoxon signed-rank test [97] is a paired statistical hypothesis test that is the nonparametric version of the paired T-test [98], which makes it suitable for small population sizes. Table 5.3 shows the p-values corresponding to the Wilcoxon sign-rank test performed on the individual sequences as well as the ALL sequence MRI tumor diagnosis tasks on the primary test set. The results are considered to be statistically significant for $p < 0.05$.

Table 5.3: Evaluating the statistical significance of the performance of the proposed model ($G(\mathcal{I})+\mathcal{I}$) over the baseline inferior modality model ($\mathcal{I}$) for the RadPath dataset. The improvements are considered to be statistically significant for $p < 0.05$. * indicates the tasks where improvements are statistically significant.

Task	p-value
T1	2.15×10^{-2} *
T2	1.38×10^{-2} *
T1Gd	3.17×10^{-1}
T2-FLAIR	2.31×10^{-2} *
ALL	6.33×10^{-2}

5.2 Derm7pt

Table 5.4 presents the results of our experiments on the Derm7pt dataset. The inferior modality of the clinical images is denoted by C and the superior dermoscopic modality is denoted by D. As in the case of RadPath, we denote the different models based on the modalities used for the task. The multimodal model is denoted by $D + C$, the modality-specific models of dermoscopic and clinical images respectively by D and C, and the guided models by $G(C)$ and $G(C) + C$. Further, similar to equation 5.1, Δ_C denotes the improvement in the performance of the guided model, $G(C) + C$ over the clinical model, C, and is computed as follows:

$$\Delta_C = \frac{\mathcal{X}_{G(C)+C} - \mathcal{X}_C}{\mathcal{X}_C} \times 100. \tag{5.3}$$

As discussed in section 4.2.2, we use the pre-trained models provided by Kawahara *et al.* [15] for both the modality-specific models (D and C) as well as the multimodal model ($D + C$). Further, based on the predictions of the 7-point criteria, we infer the melanoma diagnosis, which is described in the following section.

5.2.1 Melanoma Inference

Similar to Kawahara *et al.* [15], as we predict each of the seven 7-point criteria, the inference of melanoma can be done in two ways. The first way is to directly predict the diagnosis (DIAG), of which melanoma is a category (Table 4.2). The second way to infer melanoma is based on the 7-point criteria [83]. As mentioned in section 4.1.2, each of the 7-point criteria is assigned a score based on its association with melanoma. These scores are provided in the last column of Table 4.2.

For a given skin lesion, based on the predictions of the 7-point criteria, the 7-point score (S) can be computed by summing the scores of each of the 7-point criteria tasks, as shown in Equation 5.4. Here, the index i corresponds to the image while index j corresponds to the 7-point criteria and the $score(\cdot)$ function maps the prediction to the corresponding score. If the computed 7-point score is greater than a certain pre-specified threshold, the skin lesion would be diagnosed as melanoma positive (denoted by 1), else, negative (denoted by 0) as shown in Equation 5.5. We use the commonly used thresholds of $t = 1$ and $t = 3$ in our experiments [15, 83]. Finally, in order to compare the performance of different binary classifiers, we use the AUROC score. Table 5.5 presents the results of the melanoma inference for the aforementioned thresholds and Figure 5.3 shows the ROC curves for melanoma inference, first using direct prediction of diagnosis and second using the 7-point criteria.

$$S^i = \sum_{j=1}^{7} score(\hat{y}_j^i) \tag{5.4}$$

$$\hat{y}_{7pt}^i = \begin{cases} 1 & \text{if } S^i \geq t \\ 0 & \text{else} \end{cases} \tag{5.5}$$

Table 5.4: Derm7pt results showing the 7-point criteria and diagnosis (DIAG) predictions. Clinical (C) is guided by Dermoscopic (D). C: Inferior, D: Superior.

	Method		7-point criteria							DIAG
			PN	BWV	VS	PIG	STR	DaG	RS	
1.	$D + C$ [15]	BA↑	0.6868	0.7721	0.4728	0.5206	0.6111	0.5979	0.7311	0.4841
		F1↑	0.6939	0.8826	0.8163	0.6301	0.7347	0.6173	0.7832	0.6888
2.	D [15]	BA↑	0.6661	0.8098	0.5526	0.5735	0.6217	0.5839	0.7193	0.6350
		F1↑	0.6887	0.8596	0.7908	0.6326	0.7168	0.5969	0.7704	0.7168
3.	C [15]	BA↑	0.5851	0.6905	0.5132	0.4847	0.5815	0.5303	0.6871	0.4380
		F1↑	0.5841	0.7755	0.7474	0.5714	0.6250	0.5280	0.7142	0.6045
4.	$G(C)$	BA↑	0.5633 ±0.0143	0.6642 ±0.0108	0.3951 ±0.0058	0.4609 ±0.0089	0.5099 ±0.0134	0.5034 ±0.0075	0.6535 ±0.0150	0.3300 ±0.0278
		F1↑	0.5889 ±0.0140	0.8228 ±0.0109	0.8032 ±0.0031	0.6116 ±0.0083	0.6815 ±0.0084	0.5017 ±0.0063	0.7637 ±0.0067	0.6167 ±0.0126
5.	$G(C) + C$	BA↑	0.5919 ±0.0001	0.6965 ±0.0030	0.5282 ±0.0108	0.5023 ±0.0049	0.5488 ±0.0049	0.5401 ±0.0034	0.6959 ±0.0039	0.4471 ±0.0042
		F1↑	0.6131 ±0.0024	0.8104 ±0.0024	0.7491 ±0.0086	0.5952 ±0.0031	0.6301 ±0.0055	0.5527 ±0.0066	0.7449 ±0.0127	0.6182 ±0.0031
6.	Δ_C (%)	BA↑	+1.0	+0.8	+2.9	+3.7	(−5.6)	+1.8	+1.1	+2.0
		F1↑	+4.9	+4.5	+0.2	+4.2	+0.8	+4.5	+4.2	+2.3

Table 5.5: Melanoma Inference from the 7-point criteria predictions of Table 5.4 computed for thresholds $t = 1, 3$ alongside the AUROC score.

Method	$t = 1$		$t = 3$		AUROC
	BA↑	F1↑	BA↑	F1↑	
$D + C$	0.6734	0.5918	0.7165	0.7882	0.7883
D	0.6568	0.5867	0.7026	0.7627	0.7645
C	0.6631	0.5765	0.6915	0.7168	0.7390
$G(C)$	0.6754 ±0.0084	0.6821 ±0.0110	0.6308 ±0.0094	0.7448 ±0.0116	0.7167 ±0.0035
$G(C) + C$	0.6918 ±0.0043	0.6290 ±0.0030	0.7045 ±0.0103	0.7361 ±0.0049	0.7515 ±0.0021
Δ_C (%)	+4.2	+9.2	+1.8	+2.7	+1.6

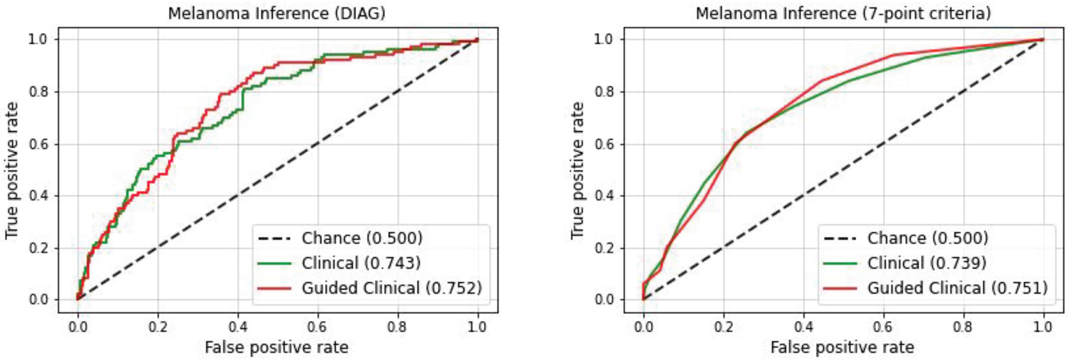


Figure 5.3: ROC curves for Melanoma Inference from the direct prediction of diagnosis (DIAG) (*left*) and using the 7-point criteria (*right*). The number in the parenthesis corresponds to the AUROC score of the classifier.

Table 5.6: Confusion matrices of the 7-point criteria and diagnosis predictions of the inferior modality model ($\mathcal{I}$) and the proposed guided model ($G(\mathcal{I}) + \mathcal{I}$) for the Derm7pt dataset. The arrows (↑, ↓) indicate the direction of improvement in the balanced accuracy scores.

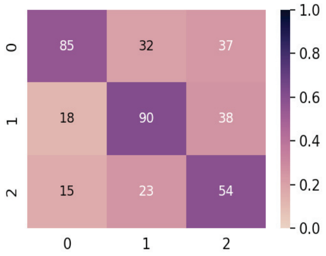
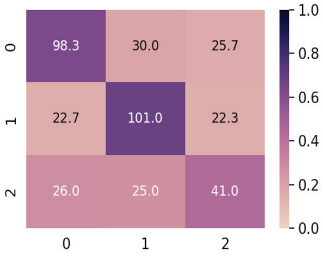
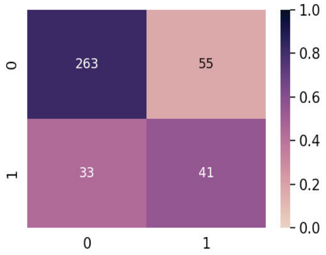
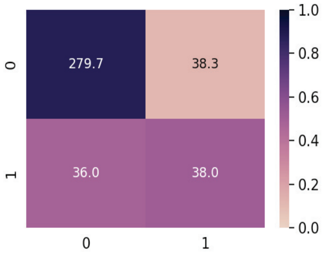
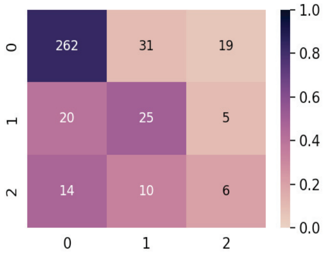
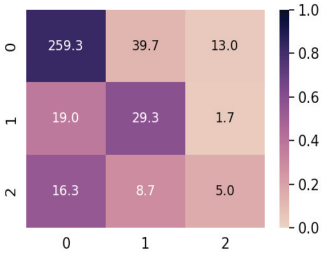
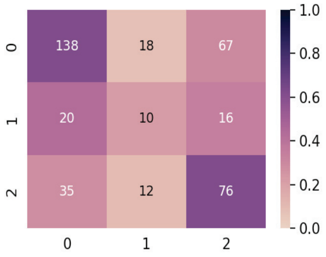
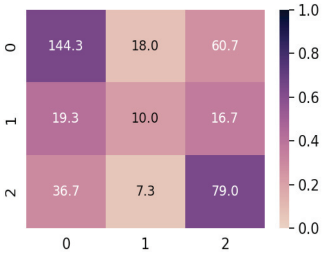
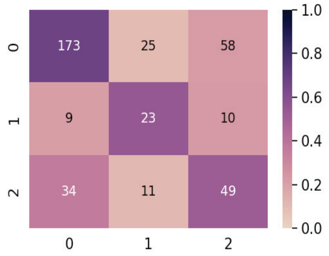
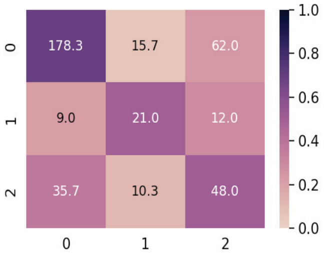
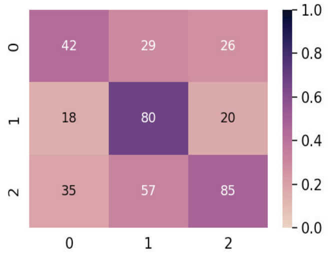
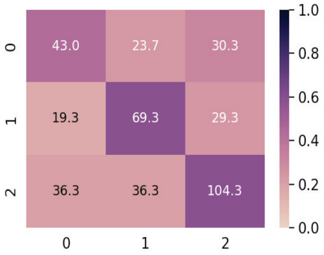
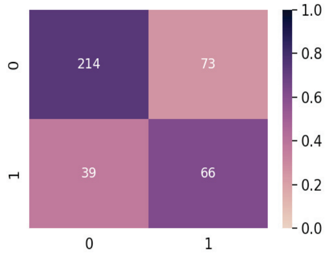
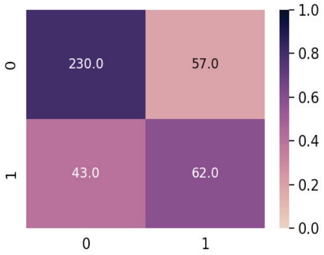
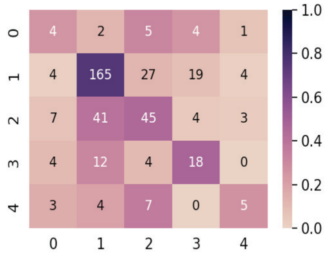
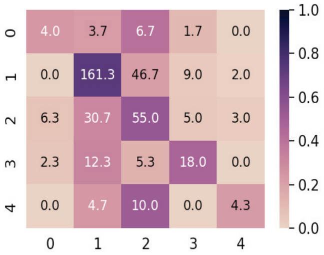
Task	$\mathcal{I}$	$G(\mathcal{I}) + \mathcal{I}$
PN(↑)		
BWV(↑)		
VS(↑)		
PIG(↑)		

Table 5.7: Confusion matrices of the 7-point criteria and diagnosis predictions of the inferior modality model ($\mathcal{I}$) and the proposed guided model ($G(\mathcal{I}) + \mathcal{I}$) for the Derm7pt dataset. The arrows (↑, ↓) indicate the direction of improvement in the balanced accuracy scores. (*Continued*)

Method	w/o guidance	w/ guidance
STR(↓)		
DaG(↑)		
RS(↑)		
DIAG(↑)		

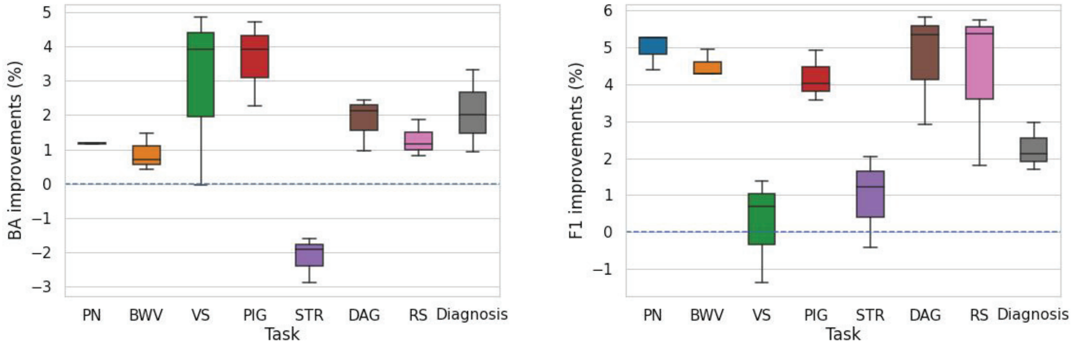


Figure 5.4: Improvements in the balanced accuracy (*left*) and F1 (*right*) scores of the guided model ($G(\mathcal{I})+\mathcal{I}$) over the baseline model (I) for the 7-point criteria and Diagnosis predictions.

5.2.2 Statistical significance

In order to evaluate the statistical significance of the performance of the proposed method ($G(\mathcal{I})+\mathcal{I}$) over the baseline inferior modality model ($\mathcal{I}$) of the Derm7pt dataset, similar to Abhishek *et al.* [99], we use the McNemar's test. McNemar's test [100] is a paired nonparametric statistical hypothesis test commonly used in DL settings where training multiple models can be expensive. Table 5.8 shows the p-values corresponding to McNemar's test performed on the 7-point criteria and diagnosis prediction tasks on the primary test set and the results are considered to be statistically significant for $p < 0.05$.

Table 5.8: Evaluating the statistical significance of the performance of the proposed model ($G(\mathcal{I})+\mathcal{I}$) over the baseline inferior modality model ($\mathcal{I}$) for the Derm7pt dataset. The improvements are considered to be statistically significant for $p < 0.05$. * indicates the tasks where improvements are statistically significant.

Task	p-value
PN	1.55×10^{-1}
BWV	2.94×10^{-2} *
VS	5.71×10^{-1}
PIG	1.17×10^{-1}
STR	6.43×10^{-1}
DaG	1.55×10^{-1}
RS	1.95×10^{-2} *
DIAG	4.57×10^{-1}

5.3 Discussion

Based on the results presented in Tables 5.1, 5.4, and 5.5, across the RadPath and Derm7pt datasets, we note the following observations. For the sake of clarity, $\mathcal{S}$ and $\mathcal{I}$ denote the superior and inferior modalities, which in the case of RadPath (Table 5.1) are denoted by P and R respectively, while in the case of Derm7pt (Table 5.4 and 5.5) are denoted by D and C respectively. In both sets of experiments, we observe the following behaviors in a majority of cases.

5.3.1 $\mathcal{S}$ outperforms $\mathcal{I}$

The superior modality outperforms the inferior modality. While it is common knowledge in the clinical setting that histopathology is more informative than radiology for tumor diagnosis, and the dermoscopic images provide a more detailed picture of the skin lesion than the clinical counterpart, the DL models from our experiments follow the same behavior. The results from the baseline models of both experiments confirm that the superior modality is more accurate for disease diagnosis. Table 5.1, shows that the classifier P significantly outperforms the individual MRI classifiers while being marginally better than the ALL sequence MRI classifier. Similarly, from Table 5.4, the classifier D outperforms the classifier C across all the 7-point criteria, diagnosis, and the overall melanoma inference in Table 5.5.

5.3.2 $\mathcal{S} + \mathcal{I}$ outperforms $\mathcal{S}$

The multimodal model outperforms the superior modality classifier. In a clinical setting, the clinician leverages all available modalities to gather complementary information before finalizing a diagnosis decision. However, in the DL setting, we observe mixed results. In the case of the RadPath dataset (Table 5.1), the multimodal model outperforms the superior modality histopathology model, thereby affirming the value added by the inferior radiology data such as the location and size of the tumor in the DL-based diagnosis. However, in the case of the Derm7pt dataset (Table 5.4 and 5.5), we observe that the multimodal model does not concretely outperform the superior dermoscopic model. The multimodal model outperforms the superior modality model in only three of the seven 7-point criteria. This behavior, as also shown earlier in the works of Abhishek *et al.* [99] and Kawahara *et al.* [15], can be attributed to the redundancy of the multimodal information in a DL setup caused due to the addition of the inferior clinical modality. The redundancy of the multimodal information that hinders the performance of the multimodal model underscores the need to develop methods such as the work proposed by Braman *et al.* [39] that orthogonalize the latent representations of different modalities so as to minimize the information redundancy.

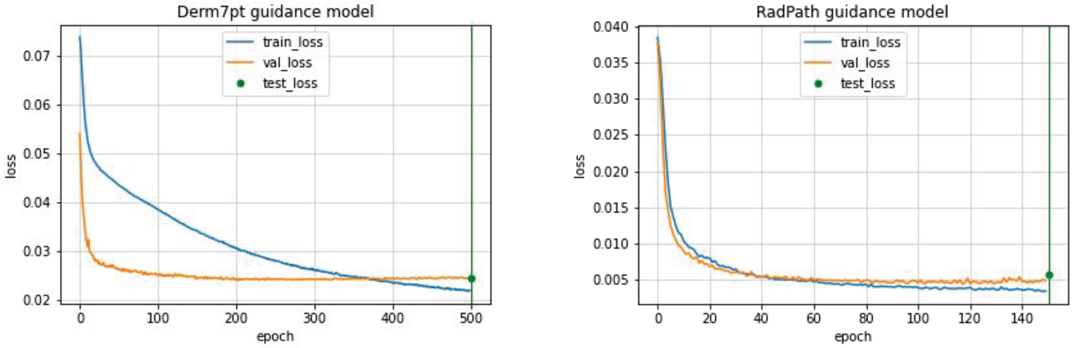


Figure 5.5: Loss curves of the guidance models of Derm7pt (*left*) and RadPath (*right*) datasets.

5.3.3 $G(\mathcal{I})$ alone does not outperform $\mathcal{I}$

Our (first) guided model in Figure 3.4 does not outperform the inferior modality classifier. As discussed in section 3.2.3, this model uses the trained guidance model as a bridge between the inferior modality latent representations and the superior modality decoder. Essentially, the performance of this model is determined by the performance of the guidance model. We observe that in both datasets, this guided model performs worse than the baseline inferior modality model. While the guidance model maps the latent representations from the inferior to the superior modality with near to zero errors as can be seen from the loss curves in Figure 5.5, the poor performance of this model could be attributed to the sensitive nature of the DL models, where a small change in the activation at an intermediate layer can change the classification output. A larger multimodal dataset can further lessen the reconstruction error in the guidance model and improve the performance of this model but we believe that fine-tuning this model using the cross-entropy loss could significantly mitigate this issue.

5.3.4 $G(\mathcal{I}) + \mathcal{I}$ outperforms $\mathcal{I}$

Our proposed model in Figure 3.5 outperforms the inferior modality classifier. As shown in this figure, this model concatenates the inferior modality latent representations with the guided representations and trains the combined decoder on the concatenated representation. The improvement in the performance of this model over the inferior modality model can be seen in Figures 5.2 and 5.4 and the ROC curves in Figure 5.3, where the melanoma inference improves when directly predicting the diagnosis as well as from the 7-point criteria. The proposed model outperforms the inferior modality model across all the individual MRI sequences as well as the ALL sequence MRI, shown in Table 5.1. Similarly, in the case of the Derm7pt dataset, we observe improvements in six of the seven 7-point criteria and diagnosis predictions, shown in Table 5.4. With this, we show that our proposed model successfully leverages the existing multimodal data to improve the performance of the novel inferior modality data.

Further, from the Figures 5.2 and 5.4, we observe that in the case of RadPath dataset, the improvements of the proposed method are prominent on the BA metric while the improvements in the case of the Derm7pt dataset are more evident on the F1 metric. This reversal of trend is mainly due to the difference in the sizes of the two datasets. The confusion matrices in Table 5.2 show that the proposed method in the case of RadPath dataset improves the performance in the case of minority classes resulting in a higher BA, while the overall performance (F1) of the proposed method is not significantly higher than the baseline method. However, the confusion matrices in Tables 5.6 and 5.7 show that in the case of Derm7pt dataset, the proposed method improves the overall performance with comparatively smaller improvements in the performance of the minority classes.

5.3.5 $G(\mathcal{I}) + \mathcal{I}$ is comparable to $\mathcal{S} + \mathcal{I}$ for RadPath

In the case of the RadPath dataset, the proposed model (Figure 3.5) performs comparably to the multimodal model (Figure 5.1) that uses both superior and inferior modalities during the inference. As shown in Table 5.1, the proposed method with only the inferior modality (ALL sequence MRI) is able to close in on the performance of the multimodal model. Additionally, it outperforms the superior modality histopathology model, which in a clinical setting would imply an alleviated need for the acquisition of the superior modality. However, we do not observe this behavior in the case of the Derm7pt dataset (Tables 5.4 and 5.5). The proposed model, while outperforming the inferior modality model, does not match the performance of the superior modality model, or the multimodal model. This can be explained by the loss curves of the two guidance models shown in Figure 5.5. The latent representations of radiology are able to map more closely to superior histopathology representations than the clinical mapping to the dermoscopic representations. This behavior can be attributed to the relative importance of the two modalities used in the diagnosis – radiology being more important to tumor diagnosis than clinical images for lesion diagnosis. Further, we hypothesize that the performance of the multimodal model ($\mathcal{S}+\mathcal{I}$) forms an upper bound on the performance achievable by the proposed model ($G(\mathcal{I}) + \mathcal{I}$) as the guidance ($G(\mathcal{I})$) only aims to mimic the latent representation of the superior modality ($\mathcal{S}$), without completely replicating it.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Motivated by the observation that, for a particular clinical task, better-performing medical imaging modalities are typically less feasible to acquire, in this work, we proposed a student-teacher method that distills knowledge learned from a better-performing (superior) modality to guide a more feasible yet under-performing (inferior) modality and steer it towards improved performance. Our evaluation on two multimodal medical imaging-based diagnosis tasks (skin and brain cancer diagnosis) demonstrated the ability of our method to boost the classification performance when only the inferior modality is used as input. We even observed (for the brain tumor classification task) that our proposed model, using guided unimodal data, achieved results comparable to a model that uses both superior and inferior multimodal data, i.e. potentially alleviating the need for a more expensive or invasive acquisition.

6.2 Limitations and Future Work

While the proposed method is successfully able to leverage the existing multimodal medical images to improve the diagnosis using the novel medical images of the inferior modality, we note that the proposed method has certain limitations and suggest directions for future work to mitigate these limitations and improve the performance of the guidance.

1. The training strategy of the proposed method involves three sequential steps (as discussed in Chapter 3) – training the modality-specific classifiers, training the guidance model, and finally, training the combined decoder. We believe that training our models in an end-to-end fashion, in a single optimization step, as opposed to multiple sequential optimizations, would enhance the performance of the proposed method. Such end-to-end trained models would result in a guidance model that optimizes the latent representations for the end goal of classification as opposed to the proposed work where they are constrained to mimic the superior modality latent representations that

might not be optimal for the classification. While this could be computationally challenging for datasets with huge images, such as RadPath, it could prove to be beneficial for datasets like Derm7pt with regular image sizes.

2. Data augmentation is a popular technique used to mitigate model overfitting and has, in turn, shown to improve model performance. In training the guidance model (as discussed in section 3.2.2), we make use of an autoencoder-like model to map from the latent representation of the inferior modality to that of the paired superior modality. The current implementation involves a one-to-one mapping from an unaugmented inferior modality latent space to the unaugmented superior modality latent space. The lack of data augmentation limits the performance of the guidance, as seen in the case of the Derm7pt dataset in Figure 5.5. However, we believe it is possible to circumvent this limitation with the use of an augmentation-invariant mapping, where the latent representations of different augmentations of an inferior modality image are encouraged to be identical before mapping it to the corresponding superior modality representation – a many-to-one mapping.

3. The surging interest in multimodal data analysis has given rise to a variety of techniques for aggregating information from multiple modalities. In this work, we use a simple concatenation of the latent representations to aggregate the multimodal information (Figure 3.5). We believe the use of advanced aggregation strategies such as (i) the use of attention before aggregation to assign due importance to the different latent representations as in the work proposed by Braman *et al.* [39] and (ii) Kronecker product-based aggregation to produce more expressive combined representations as in the work proposed by Chen *et al.* [38] could improve the performance of the proposed method.

4. While the proposed method successfully improves the diagnosis performance in the case of 2D images of skin lesions (Derm7pt) and 3D images of multi-sequence MRIs (RadPath), given the variety of modalities involved in the acquisition of medical data, such as the tabular metadata, sequential genomic data, etc., the future work could leverage such non-imaging modalities in the guidance of the inferior modality. The only difference in this case would be the replacement of the CNN-based encoder with an appropriate modality-specific encoder.

5. As previously mentioned, our work uses multimodal medical images to improve the diagnosis using the novel images of the inferior modality. However, given the prevalent issue of DL models failing on out-of-distribution data, the novel images of the inferior modality have to be acquired from the same scanner while following the same acquisition protocols. The difference in acquisition protocols between different hospitals significantly limits the applicability of our method. Future works should be able to

leverage the multimodal data of Hospital A to improve the diagnosis done at Hospital B. In order to achieve this, the current work needs to be extended in the direction of domain generalization to handle the distribution shift in data.

6. One of the limitations of the proposed method that hinders its widespread applicability is the requirement of paired multimodal datasets. Unpaired multimodal KD has been employed by Dou *et al.* [78] and Li *et al.* [79] for medical image segmentation, where they leverage the anatomical similarity along with the co-registered images across different modalities. The anatomical structures more evident in the teacher modality are used to improve the segmentation of the same structures in the student modality. Future works along the lines of unpaired multimodal guidance for classification tasks would significantly improve the utility of the proposed work.

Bibliography

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[2] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[3] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[4] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[5] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[7] Yiping Shao, Simon R Cherry, Keyvan Farahani, Ken Meadors, Stefan Siegel, Robert W Silverman, and Paul K Marsden. Simultaneous pet and mr imaging. *Physics in Medicine & Biology*, 42(10):1965, 1997.

[8] Thomas Beyer, David W Townsend, Tony Brun, Paul E Kinahan, Martin Charron, Raymond Roddy, Jeff Jerin, John Young, Larry Byars, and Ronald Nutt. A combined pet/ct scanner for clinical oncology. *Journal of nuclear medicine*, 41(8):1369–1379, 2000.

[9] Kenneth Aldape, Gelareh Zadeh, Sheila Mansouri, Guido Reifenberger, and Andreas von Deimling. Glioblastoma: pathology, molecular mechanisms and markers. *Acta neuropathologica*, 129(6):829–848, 2015.

[10] Adriana Olar and Kenneth D Aldape. Using the molecular classification of glioblastoma to inform personalized treatment. *The Journal of pathology*, 232(2):165–177, 2014.

[11] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of

tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[12] Yoo-Ah Kim, Dong-Yeon Cho, and Teresa M Przytycka. Understanding genotype-phenotype effects in cancer via network approaches. *PLoS computational biology*, 12(3):e1004747, 2016.

[13] Clipart library. `http://clipart-library.com/clipart/265773.htm`. [Online; accessed 10-May-2022].

[14] Tahsin Kurc, Spyridon Bakas, Xuhua Ren, Aditya Bagari, Alexandre Momeni, Yue Huang, Lichi Zhang, Ashish Kumar, Marc Thibault, Qi Qi, et al. Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches. *Frontiers in neuroscience*, page 27, 2020.

[15] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[16] Consumer Reports. The surprising dangers of ct scans and x-rays. `https://www.consumerreports.org/cro/magazine/2015/01/the-surprising-dangers-of-ct-sans-and-x-rays/index.htm`, 2015. Accessed: 2022-04-04.

[17] Alan S Brody, Donald P Frush, Walter Huda, Robert L Brent, and Section on Radiology. Radiation risk to children from computed tomography. *Pediatrics*, 120(3):677–682, 2007.

[18] Madan M Rehani, Kai Yang, Emily R Melick, John Heil, Dušan Šalát, William F Sensakovic, and Bob Liu. Patients undergoing recurrent ct scans: assessing the magnitude. *European radiology*, 30(4):1828–1836, 2020.

[19] Stacy Loeb, Annelies Vellekoop, Hashim U Ahmed, James Catto, Mark Emberton, Robert Nam, Derek J Rosario, Vincenzo Scattoni, and Yair Lotan. Systematic review of complications of prostate biopsy. *European urology*, 64(6):876–892, 2013.

[20] Quinn T Ostrom, Haley Gittleman, Gabrielle Truitt, Alexander Boscia, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2011–2015. *Neuro-oncology*, 20(suppl_4):iv1–iv86, 2018.

[21] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[22] Quinn T Ostrom, Luc Bauchet, Faith G Davis, Isabelle Deltour, James L Fisher, Chelsea Eastman Langer, Melike Pekmezci, Judith A Schwartzbaum, Michelle C Turner, Kyle M Walsh, et al. The epidemiology of glioma in adults: a "state of the science" review. *Neuro-oncology*, 16(7):896–913, 2014.

[23] Samanta Vicente Oliveira, Andre Caroli Rocha, Marcelo Minharro Ceccheti, Camila de Barros Gallo, and Fábio Abreu Alves. Odontogenic myxoma in a child treated with enucleation and curettage. *Autopsy & case reports*, 8(3), 2018.

[24] Michael T McCann, John A Ozolek, Carlos A Castro, Bahram Parvin, and Jelena Kovacevic. Automated histology analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine*, 32(1):78–87, 2014.

[25] Aïcha BenTaieb and Ghassan Hamarneh. Deep learning models for digital pathology. *arXiv preprint arXiv:1910.12329*, 2019.

[26] Wikipedia. Magnetic resonance imaging — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Magnetic%20resonance%20imaging&oldid=1077305794`, 2022. [Online; accessed 12-April-2022].

[27] Catharine Paddock. Brain scans could help predict whether antidepressants will work. `https://www.medicalnewstoday.com/articles/326510`, Sep 2019. [Online; accessed 16-Apr-2022].

[28] World Health Organization. Radiation: Ultraviolet (uv) radiation and skin cancer. `https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer`, Oct 2017. [Online; accessed 13-April-2022].

[29] American Cancer Society. Key statistics for melanoma skin cancer. `https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html`, Jan 2022. [Online; accessed 14-Apr-2022].

[30] Frederick C Beddingfield III. The melanoma epidemic: res ipsa loquitur. *The Oncologist*, 8(5):459–465, 2003.

[31] Giuseppe Argenziano and H Peter Soyer. Dermoscopy of pigmented skin lesions–a valuable tool for early. *The lancet oncology*, 2(7):443–449, 2001.

[32] Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

[33] Amanda Oakley. Dermoscopy. `https://dermnetnz.org/topics/dermoscopy`, 2004. [Online; accessed 16-Apr-2022].

[34] Mayur Mallya and Ghassan Hamarneh. Deep multimodal guidance for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022.

[35] Mayur Mallya and Ghassan Hamarneh. Deep multimodal guidance for medical image classification. *arXiv preprint arXiv:2203.05683*, 2022.

[36] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review. *arXiv preprint arXiv:2203.15588*, 2022.

[37] Yan Xu. Deep learning in multimodal medical image analysis. In *International Conference on Health Information Science*, pages 193–200. Springer, 2019.

[38] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020.

[39] Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, Caleb Willis, Martin C Stumpe, and Jagadish Venkataraman. Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 667–677. Springer, 2021.

[40] Luis Gómez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

[41] Linmin Pei, Lasitha Vidyaratne, Md Monibor Rahman, and Khan M Iftekharuddin. Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images. *Scientific Reports*, 10(1):1–11, 2020.

[42] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 978–987, 2016.

[43] Edgar A Bernal, Xitong Yang, Qun Li, Jayant Kumar, Sriganesh Madhvanath, Palghat Ramesh, and Raja Bala. Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. *IEEE Transactions on Multimedia*, 20(1):107–118, 2017.

[44] Héctor P Martínez and Georgios N Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*, pages 34–41, 2014.

[45] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019.

[46] Ashnil Kumar, Jinman Kim, David Lyndon, Michael Fulham, and Dagan Feng. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics*, 21(1):31–40, 2016.

[47] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[48] Baocai Yin, Hu Cheng, Fengyan Wang, and Zengfu Wang. Brain tumor classification based on mri images and noise reduced pathology images. In *International MICCAI Brainlesion Workshop*, pages 465–474. Springer, 2020.

[49] Xiao Ma and Fucang Jia. Brain tumor classification with multimodal mr and pathology images. In *International MICCAI Brainlesion Workshop*, pages 343–352. Springer, 2019.

[50] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[51] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[52] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016.

[53] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[54] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016.

[55] Yue Ruan, Han-Hung Lee, Ke Zhang, and Angel X Chang. Tricolo: Trimodal contrastive loss for fine-grained text to shape retrieval. *arXiv preprint arXiv:2201.07366*, 2022.

[56] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[57] Seungwhan Moon, Suyoun Kim, and Haohan Wang. Multimodal transfer deep learning with applications in audio-visual recognition. *arXiv preprint arXiv:1412.3121*, 2014.

[58] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

[59] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[60] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[61] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[62] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[64] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514–63537, 2020.

[65] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of applied clinical medical physics*, 22(1):11–36, 2021.

[66] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[67] Vasant Kearney, Benjamin P Ziemer, Alan Perry, Tianqi Wang, Jason W Chan, Lijun Ma, Olivier Morin, Sue S Yom, and Timothy D Solberg. Attention-aware discrimination for mr-to-ct image translation using cycle-consistent generative adversarial networks. *Radiology: Artificial Intelligence*, 2(2):e190027, 2020.

[68] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39(4):1170–1183, 2019.

[69] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[70] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[71] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision. *Neurocomputing*, 454:25–33, 2021.

[72] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019.

[73] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[74] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[75] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020.

[76] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018.

[77] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018.

[78] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.

[79] Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 775–783, 2020.

[80] Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 772–781. Springer, 2020.

[81] Tom van Sonsbeek, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational knowledge distillation for disease classification in chest x-rays. In *International Conference on Information Processing in Medical Imaging*, pages 334–345. Springer, 2021.

[82] Alessandro Crimi and Spyridon Bakas. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I*, volume 12658. Springer Nature, 2021.

[83] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12):1563–1570, 1998.

[84] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[85] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[86] Marvin Lerousseau, Eric Deutsch, and Nikos Paragios. Multimodal brain tumor classification. In *International MICCAI Brainlesion Workshop*, pages 475–486. Springer, 2020.

[87] Azam Hamidinekoo, Tomasz Pieciak, Maryam Afzali, Otar Akanyeti, and Yinyin Yuan. Glioma classification using multimodal radiology and histology data. In *International MICCAI Brainlesion Workshop*, pages 508–518. Springer, 2020.

[88] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

[90] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[91] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[92] MONAI Consortium. MONAI: Medical Open Network for AI. `https://github.com/Project-MONAI/MONAI`, 2 2022.

[93] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[94] Compute canada. `https://www.computecanada.ca/`. [Online; accessed 09-August-2022].

[95] Rosario Delgado and Xavier-Andoni Tibau. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916, 2019.

[96] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[97] Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.

[98] Scipy documentation. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html`. [Online; accessed 12-August-2022].

[99] Kumar Abhishek, Jeremy Kawahara, and Ghassan Hamarneh. Predicting the clinical management of skin lesions using deep learning. *Scientific reports*, 11(1):1–14, 2021.

[100] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.