Deep Learning Techniques in MRI images for Brain Cancer Segmentation and Detection through Domain Adaptation via Dataset Fusion and Semi Supervised Learning strategies

LIEW MEI WERN ANDREA

M.Eng

A thesis submitted for the degree of Doctor of Philosophy at

Monash University in 2023

School of Engineering

New page

1 Copyright notice

New page

2 Abstract

Primary and secondary brain tumors are fundamentally different in appearance. Consequently, a primary brain segmentation scheme trained with only primary gliomas might not be able to segment secondary brain tumors due to its multiplicity and minute size. In this thesis, standalone deep learning architectures were introduced for the segmentation of primary gliomas and detection of secondary metastases. These systems were enhanced with CNN-based blocks incorporating a-priori knowledge, morphological indicators, and expert soft knowledge. The work in chapters 4,5,6 will describe the concepts behind these enhancements to provide saliency and attention to the deep learning architectures.

This thesis expands from work carried out on a publicly available primary brain tumor dataset and ultimately led to the development of the Sunway Brain Metastases Dataset (SUNMET), which contains 121 patients from a Malaysian demographic, which can be seen as complementary to the other public and private metastases datasets published. The most similar dataset on brain metastases from Asia is presented by Shanghai JiaoTong University which consists predominantly of patients from China. It is hoped that the findings from this thesis will assist in Computer Aided Diagnosis for the clinical radiologists at Sunway Medical Center for early intervention, as the outcomes of Gamma Knife Therapy on smaller metastases are more favorable. Therefore, it is imperative that no metastases are excluded when quantifying performance measures.

The work in this thesis addresses the deep learning architectures necessary for the segmentation and detection tasks, as well as the training strategies that are used to improve the model performance. Three architectures were purpose-built for this, namely, CASPIANET++, NLMET and STROLLNET. Accompanying each architecture is a training paradigm which addresses one of the most common problems in deep learning of medical images, data scarcity.

In segmenting primary tumors, CASPIANET++ features multidimensional channel-spatial blocks, which incorporate a-priori morphological indicators as attention blocks. The resultant architecture trained on unlabeled data from the dataset through the Noisy Student Curriculum Learning strategy showed improvement over state-of-the-art models with higher training capacities.

Next, NLMET was developed to detect secondary tumors with sizes ranging from 2-40mm. A further two datasets were enlisted for training the detection module, as only 98 data points were available in SUNMET. To bridge the domain gaps between the secondary and primary datasets, a CAVEAT learning scheme gradually introduces higher confidence level of pseudo labels and larger number of samples according to lesion volume at every self-training cycle. While NLMET with CAVEAT performed well, an opportunity for optimizing the non-local blocks by incorporating soft knowledge of the expert annotator was presented. STROLLNET uses short-range non-local blocks that act within the spatial view of a typical metastatic lesion, as opposed to its entire volume. This new generation STROLLNET reduced the FP/patient and increased the Sensitivity of detection compared to its NLMET counterpart. The PHYSCISION approach to dataset fusion with a breast cancer dataset was implemented, which is advantageous as it eliminates the uncertainty of pseudo labels, requiring only a single cycle of training and underscores the potential for using dissimilar medical datasets in training.

New page (standard thesis only)

3 Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: ……………………

Print Name: Liew Mei Wern Andrea

Date: 13/3/2023

4 Publications during enrolment

This thesis contains material from 3 papers published/submitted as follows where I was the first author.

Andrea Liew, Chun Cheng Lee, Boon Leong Lan, Maxine Tan, (2021),

CASPIANET++: A multidimensional Channel-Spatial Asymmetric attention network with Noisy Student Curriculum Learning paradigm for brain tumor segmentation, Computers in Biology and Medicine, Volume 136, 104690, ISSN 0010-4825, (published)

Andrea Liew, Chun Cheng Lee, Valarmathy Subramaniam, Boon Leong Lan, Maxine Tan (2023), Gradual Self-Training via Confidence and Volume Based Domain Adaptation for Multi Dataset Deep Learning-Based Brain Metastases Detection Using Nonlocal Networks on MRI Images. J Magn Reson Imaging. https://doi.org/10.1002/jmri.28456 (published)

Andrea Liew; Chun Cheng Lee; Hong Hui Yeoh; Boon Leong Lan; Maxine Tan (2023),

Improving Brain Metastases Detection in T1-MPRAGE scans through Modelling Short-Range Dependencies and Dataset Fusion Strategies. European Radiology (submitted)

List of publications where I was co-author:

Hong Hui Yeoh; **Andrea Liew**; Raphaël Phan; Fredrik Strand; Kartini Rahmat; Tuong L Nguyen; Jennifer Stone; John L. Hopper; Maxine Tan, RADIFUSION: A multi-radiomics deep learning based Breast cancer risk Prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement. Medical Image Analysis (submitted)

List of conference proceedings published in an international conference:

Andrea Liew; Chun Cheng Lee; Boon Leong Lan; Maxine Tan (2022), A novel curriculum based paradigm for self-training deep learning networks for brain glioma segmentation. CARS 2022—Computer Assisted Radiology and Surgery Proceedings of the 36th International Congress and Exhibition Tokyo, Japan, June 7–11, 2022. Int J CARS 17 (Suppl 1), 1–147 (2022). https://doi.org/10.1007/s11548-022-02635-x

Andrea Liew; Chun Cheng Lee; Boon Leong Lan; Maxine Tan (2023), Creating Pseudo Datasets from Disparate Datasets for Domain Adaptation for Brain Metastases Detection. Accepted for publication in CARS 2023 (June 20-23, 2023)

New page

5 Acknowledgements

This research was supported by the Graduate Research Merit Scholarship awarded by Monash University and a grant from Sunway Medical Center. I would also like to express my gratitude to Dr Maxine Tan for her invaluable insights, experience, and unwavering support throughout my candidature. I also wish to thank Dr Lee Chun Cheng and Mr Liew Tuck Fook from the Radiology Department, Sunway Medical Center, for all your support in procuring the Sunway Medical Center Brain Metastases dataset. Lastly, I thank George Liew, Nicholas Liew, Ong Chuin Liu, Catherine Lim, Joni Chan for their technical, moral, and monetary support.

In loving memory, Susan Lim & Elisa Powell

6 Table of Contents

7 List of Figures

8 List of Tables

9 List of abbreviations

2D - 2 Dimensional

3D - 3 Dimensional

BD - Breast Density

BM -Brain Metastases

BraTS - Brain Tumor Image Segmentation Dataset

CT – Computed Tomography

CAD – Computer Aided Diagnosis

CNN – Convolutional Neural Networks

CASPIAN - Channel and Spatial Asymmetric Attention

CAVEAT - Confidence and Volume Based Domain Adaptation

DMF – Dilated Multi-Fiber

ET - Enhancing Tumor

FP – False positives

FP/patient - FP rate per patient

FROC - Free receiver operating characteristic curves

GPU – Graphical Processing Unit

HGG - High Grade Glioma

HFS - head-first supine

LGG - Low Grade Glioma

MRI – Magnetic Resonance Imaging

MT - Molecular type

NLB - Non-local block

NSCL – Noisy Student Curriculum Learning

NLAM – Non-local attention maps

PHYSCISION - Physical attributes and clinical factors data fusion

RPN - Region Proposal Networks

SRS - Stereotactic radiosurgery

STROLL - Short-Range dependencies for non-local

SUNMET - Sunway Medical Center Brain Metastases dataset

T1-MPRAGE – T1 Magnetization Prepared - Rapid Gradient Echo

TC - Tumor Core

WT - Whole Tumor

1 Introduction

1.1 Motivation

Brain and nervous system tumors are the primary cause of mortality in young adolescents below the age of 20 [1]. In 2022, it is estimated that 18280 deaths will result from brain tumors in the United States alone. Brain cancers are classified into primary and secondary, depending on the origin of the cancerous cells. Treatment of brain tumors typically consists of irradiating the tumor bed while exercising caution in avoiding healthy cells in its proximity. To minimize the impact of radiation on healthy cells, accurate determination of tumor volumes is paramount [2]. Conventionally, oncologists manually register the gross tumor volume by outlining 2D images one by one (this could be in the hundreds for a single patient) [3]. This process is repetitive, time-consuming, and prone to human error. Additionally, researchers found significant interobserver disagreement between the location of a tumor, suggesting that this could be the reason for the poor prognosis of brain tumors [4]. To this end, an automated scheme to differentiate tumors from healthy brain cells will overcome difficulties in manual identification.

The research centers on accurate detection of primary and secondary brain tumors. We can largely breakdown our approach into three phases namely: 1) segmentation of primary gliomas, 2) detection of primary and secondary brain tumors and 3) improve accuracy in detection of primary and secondary brain tumors using out – of – domain datasets. In Phase 1, the main focus was on primary gliomas, which are the **most common primary brain cancer**. **The prognosis is bleak, where the five-year survival rate of primary gliomas is only 5.6%.** Accurate brain tumor segmentation which delineates healthy regions, and unhealthy (tumors and necrotic) regions is crucial to assist in early diagnosis, treatment, and monitoring of disease progression [5]. However, developing lesion segmentation algorithms with high accuracies to which it is clinically acceptable is not without its challenges. The heterogeneity of lesions in its appearance, locality, shape, size and number of occurrences complicate the development of segmentation rules [6]. In Phase 2 & 3, our work shifts to secondary brain metastases (BM), **the most common neurologic complication of cancer** [7]. In deep learning applications for medical images, fully annotated data is insufficient to train the network to a good accuracy. To increase data volume in the training set, out – of - domain datasets are jointly used in deep learning applications. Another source of data comes from unlabeled medical images, which are pseudo labeled via self-training paradigms.

Authors have contended that classical machine learning methods suffer from its inability to generalize as it depends on hand-crafted features [8]. Using these hand-crafted features may not hold as lesions vary in shapes and sizes from patient to patient. With the availability of affordable and powerful GPUs, interest in deep learning methods has proliferated and achieved accuracies that surpass classical machine learning methods.

To date, the failure of CAD systems to be adopted clinically is due to the copious amounts of false positives which creates overheads for radiologists [9]. Deep learning approaches have been applied by many in commercial natural image and language processing applications. However, the progress in medical image processing is slower, due to the scarcity of datasets available for training. There are only a handful of public medical imaging datasets, of which the largest contains 1200 patients compared to natural images which can contain up to 14,000,000 datapoints.

1.2 Problem Statement

This thesis looks at two primary brain MRI datasets – one is a publicly available dataset **Bra**in **T**umor Image **S**egmentation Dataset (BraTS) consisting of primary gliomas and the other is a dataset obtained from Sunway Medical Center with brain metastasis patients. This dataset shall henceforth be known as **Sun**way Brain **Met**astases Dataset (SUNMET). While these two datasets contain malignancies affecting the same anatomy, there are differences in morphology which affect how the problem is addressed. Physical traits aside, there are also technical considerations in the dataset structure.

Metastases and primary gliomas are fundamentally different because of their size and multiplicity. Metastases found in SUNMET are between 2-40mm whereas primary gliomas are around 100mm in diameter. Some metastases detection systems may ignore any lesions 10mm and under due to its poor visibility and to lower false positives. However, Gamma Knife Therapy procedures are sensitive up to 0.5mm - so detection of smaller metastases will enable treatment to be provided when lesions are smaller and less invasive. However, smaller structures prove difficult to detect because they look like blood vessels – which introduces more undesirable false positives.

In larger structures such as gliomas, where accurate identification of structures is crucial for treatment and follow-up, it is no longer possible to simply localize the lesions as done in metastases. Here, segmentation is required to delineate accurately the different regions within the primary glioma. The three regions to be identified vary based on the category of the glioma and the disease progression when the MRI was captured. On top of this being a multiclass segmentation problem, we had to contend with severe data imbalance between the edema and other tumor regions.

As with most disease or anomaly-based datasets, it is common that there are not as many samples given that a disease affects a small proportion of the human population. This is also common in credit card transactions, where there could be a handful of illicit transactions out of many legitimate ones. This presents a few problems in training a deep learning architecture. Firstly, it is widely known that deep learning system generalizes on many examples. Next, an algorithm trying to maximize correct predictions to predict the outcome of the most prevalent class – which in our case is healthy tissue.

1.3 Research Objectives

The objectives are listed below:

1. To develop a deep learning architecture capable of brain tumor segmentation

2. To investigate self-training methods to leverage unlabeled data to improve the segmentation performance

3. To prepare a fully annotated dataset based on MRI data from 172 patients from SUNMET dataset

4. To develop a system to detect metastases on SUNMET data

5. To investigate the effectiveness of combining in-domain and out-of-domain data to improve the performance of metastases detection and to improve generalizability of the model

6. To evaluate the importance of attention mechanisms towards improving deep learning architectures

1.4 Research Contributions

Our main contributions are: -

1. CASPIANET++, A morphological based multidimensional spatial and channel attention for brain tumor segmentation
2. Implemented Noisy Student Curriculum Learning paradigm to leverage unlabeled data
3. SUNMET, a fully annotated dataset of 121 patients with brain metastases
4. NLMET is a brain metastases detection architecture with non-local block support
5. CAVEAT, an adaptive approach that combines gradual domain adaptation and self-training of in-domain and out-of-domain datasets to overcome data scarcity in the SUNMET dataset
6. STROLLNET, an improvement to NLMET consisting of an architectural update to the non-local block in NLMET which utilizes short range dependencies in non-local blocks
7. PHYSCISION, a dataset fusion paradigm which harnesses out-of-domain datasets to improve the training

1.5 Thesis Organization

This dissertation will describe the deep learning architectures along with the learning paradigms used for primary glioma segmentation and secondary brain metastases detection. The thesis is organized as follows:

Chapter 2 provides an in-depth review of segmentation and detection architectures in deep learning. It consists of a description of the open-source datasets earmarked for investigation to alleviate data scarcity. This chapter also explores alternative methods to co-train these datasets from semi supervised learning to self-training. Subsequently, the discussion diverges into two separate tasks, one on primary glioma segmentation and the other on secondary metastases detection.

Chapter 3 gives an overview of all the datasets used in this thesis, along with an overview of the systems developed.

Chapter 4 discusses the CASPIANET++ architecture implementation in detail – combining spatial and channel attention with morphological indicators. In this deep learning landscape, a single dataset is used while leveraging the unlabeled data in the validation dataset for self-training. Hence, the Noisy Student Curriculum Learning paradigm was proposed.

Chapter 5 discusses the SUNMET dataset and the corresponding demographic data was presented. The architecture of the metastases detection system was then described together with CAVEAT learning approach to combine the primary glioma dataset used previously with the secondary metastases dataset.

Chapter 6 talks about further improvements to be made on the metastases detection system to reduce the number of false positives. In this chapter, a new non-local block was introduced with attempts to increase the performance of the metastases' detection network. In efforts to augment the work on domain adaptation, breast cancer datasets were incorporated to further improve the network performance.

Chapter 7 provides context to the work done in the previous three chapters and recap all the outcomes from the project. Here, the limitations of the work will be addressed together with any potential improvements.

2 Literature Review

2.1 Overview

Convolutional neural networks (CNNs), deep learning and large datasets are the trifecta in the advancement in image segmentation and detection. Deep learning networks were created on the pretext that a single CNN layer 7x7 kernel offers less information and uses more parameters than a two-layer 3x3 kernel implementation. Modern image detection architectures such as Region Proposals Networks (RPN) are based upon CNNs trained on thousands of natural images [10].

To date, CNN based methods remain state-of-the-art for medical image detection – specifically in our area of interest, brain tumor segmentation [6][11] and brain metastases detection [12]–[16]. In deep learning applications for medical images, the volume of fully annotated data is often insufficient to train the network to a good accuracy.

In this section, we detail the popular segmentation and detection architectures used for medical tasks and discuss attention mechanisms for region saliency. Next, we address techniques to alleviate data scarcity in medical images, particularly in Semi Supervised Learning techniques and Domain adaptation.

2.2 Brain Tumor Segmentation and Brain Metastases Detection

Here, we discuss two major tasks in our study, the first involving segmentation of brain tumors on the BraTS dataset, developed as an annual competition dataset. In brain tumor segmentation every pixel in an MRI scan is assigned a class label to differentiate tumor regions and its background. For the SUNMET dataset, the ground truth is given by a means of a line or RECIST definitions, which can be transformed into a detection task. Detection tasks localize the lesion using centroid coordinates and diameter and is not concerned with class of each pixel in the MRI. Though both tasks are performed using deep CNN networks, we review the algorithms available for brain tumor segmentation and brain metastases detection separately.

2.2.1 Brain tumor segmentation algorithms

In the BraTS 2013 online evaluation system, the Cascaded Two-Pathway CNN architecture outperformed state-of-the-art methods in accuracy and speed [11]. The two-pathway architecture has two kernels of different dimensions to provide both local and global context to the model. Additionally, cascading provides additional information on nearby pixels to overcome the problem of uncorrelated predictions in adjacent pixels. A similar two-pathway architecture proposed in [17] replaces one of the CNN layers in the cascade with an Inception network. Both approaches use a patch to label a single pixel, which is resource intensive and time consuming.

While patch-based architectures predicting the label of a center pixel described in [18] emerged top in the Electron Microscopy challenge in 2012, processing results was slow, and there was redundancy due to overlapping patches. Additionally, providing the network with bigger patches required more max-pooling layers, which affected the localization accuracy. Conversely, reducing patch sizes caused a reduction in contextual information. Work on cascaded hierarchical models trained at different resolutions showed that it was possible to provide the classifier with both localized and contextual information [19]. While schemes above are the most popular choices for classification tasks, they produce pixel-wise segmentation from patch-based inputs, which is computationally expensive for slice-wise MRI data and large datasets.

Popular CNN architectures such as AlexNet, VGG nets, and GoogLeNet, originally built as classifiers, were adapted to produce dense inference for semantic segmentation. CNN classifiers inherently learn objects' heatmaps, wherein by replacing the classifier layer with a convolution layer, the network can perform dense inference [20]. To achieve end-to-end semantic segmentation, researchers have drawn upon the autoencoder structure, consisting of an encoder, which extracts feature maps, and a decoder to decode information in the feature maps to produce an output segmentation [21]. The most popular end-to-end architecture is UNET, which won first place in the ISBI cell tracking challenge in 2015 [22]. The UNET architecture is significant in biomedical image segmentation and will form the base model for the work outlined in this thesis, therefore a comprehensive description of the architecture is deemed necessary.

The input image is subjected to two successive 3x3 convolution and ReLU blocks, followed by a 2x2 max pooling operation with a stride of 2 for downsampling. The downsampled images are subjected to convolutions to double their feature maps. These maps facilitate classification tasks by acting like feature detectors for the input image, with earlier layers detecting lower-level features such as edges. In comparison, deeper layers extract shapes and patterns. As layers in the contracting path go deeper and deeper, the max pooling operation downsamples the image, as the convolutions had transformed spatial information into more complex representations. In essence, it has lost data on "where" while gaining knowledge on "what" is in the image.

The expanding path of the UNET is symmetric to the contracting path. The main idea of the expanding path is to decode the feature information encoded by the contracting path. Recall that spatial information is lost during the contracting path, so the expanding path attempts to reconstruct them through up-convolutions. The reconstructed output will produce the predictions for the classification tasks. Skip connections are applied by concatenating the output of the expanding path with the output of the up-convolutions serving as auxiliary spatial information for the decoder. The concatenations are fed into two 3x3 convolutions and ReLU block (similar to the contracting path) to facilitate more precise localization of predictions.

However, the vanishing gradient problem limits the UNET architecture from having deeper layers. The FusionNet approach takes from the ResNet architecture, adding residual blocks to allow more information flow across the network and to increase network depth [23]. This method outperforms the original UNET in electron microscopy segmentation tasks. In [24], an auxiliary autoencoder network is used in conjunction with the primary network to overcome the vanishing gradient problem. The concept of autoencoder regularization is also described by [25], the winning architecture for the BraTS 2018.

Despite the shortcomings of UNET, the second runner up of the BraTS 2018 competition showed that a well-trained vanilla UNET implementation complemented with additional training data, post-processing methods, region-based training, and loss functions can still perform very well [26]. The team showed that their UNET architecture fared better than the BraTS 2018 winning architecture in segmenting enhancing tumor (ET) regions.

ET regions (denoted by the label 4) in HGG patients are particularly challenging to segment [27]. Participants across the board report lower dice coefficients for the ET region compared to the whole tumor region in the BraTS competition [25], [28], [29],[30]

Most of the deep learning models reviewed so far utilize 2D operations, which appear counterintuitive as 3D MRI brain images are processed individually and reconstructed to 3D. Deep learning applications on 2D data results in a loss of contextual representations in the 3rd dimension. Other roadblocks preventing 3D type architectures are slow inference times of 3D convolutions and limitations in GPU processing power. Hybrid versions of 2D CNN architectures termed as 2.5D methods use coronal, axial, and sagittal views as channels in a 2D CNN implementation [31]. Another 2.5D scheme described in [32] has nine paths from normalizations of coronal, axial, and sagittal planes fed into a 2D CNN architecture. The model extends into 2.5D with the introduction of a postprocessing block consisting of 3D convolutions for the ensembling of results from multi-pathways to form a 3D output for predictions. While these hybrid methods benefit from having fewer weights and is computationally more efficient and reports improvements over their 2D counterparts, they are effectively using slice-wise information and completely ignore the volumetric information contained in 3D MRI data.

DeepMedic is an 11 layer 3D Fully CNN, leveraging on dense predictions described in [21], for more efficient 3D processing. DeepMedic features a dual pathway for multi resolutions quite similar to that described in [11] with a Conditional Random Field for post-processing [6]. The winning team in BraTS 2017, was the same team that developed DeepMedic. Their winning method, Ensemble of Multiple Models and Architectures (EMMA), combined the strengths of two DeepMedics, three UNETs, and three FCNs [20]. Models were trained under different configurations and input sizes, each producing a class confidence map. The label with the highest confidence is assigned as the voxel label. BeEMMA increased the robustness of neural networks from suboptimal configurations, and from making unpredictable errors [28]. **Ensembling** also removes variances from different model architectures, hyperparameters, and configurations that inherently perform better on certain structures.

Model cascading is another intuitive approach to solving detection problems where a first layer model detects a whole tumor, and the result is used on subsequent layers to identify more intricate details of the lesions (such as the enhancing tumor region, etc.). In 3D schemes, model cascading is cumbersome and costly as it requires the use of multiple networks and training. Therefore, authors in [33] proposed the OMNet architecture, which consists of a **shared backbone model** and utilizes a curriculum learning method by training the model with the most straightforward task first.

2.2.2 Brain metastases detection

To the best of our knowledge, deep learning methods for brain metastases (BM) detection are mainly conducted on private datasets. The largest study involves 1652 patients from 3 hospitals using a cascaded CNN strategy, one for detection of regions of interest and the other for segmentation [34]. Other methods had used established architectures such as GoogLeNet [14], Faster RCNN [12] and DeepMedic [13]. Table I provides a summary of the sensitivity and FP/scan of various BM detection methods. However, we cannot adequately compare methods because the datasets are different. Secondary BMs could originate from a host of different primary cancers, which may influence the appearance and size of the BMs [35]. To highlight the diversity in these datasets, each one we reviewed contain different distributions of primary cancers dataset while the dataset from Jünger et al. contained only BM from Non-small Cell Lung Cancers (NSCLC) [10]. Recently, the team at Stanford has made the BrainMetShare dataset publicly available, which we utilize to assess the performance of our method [14].

To improve sensitivity and reduce false positive (FP) detections of BM, two-stage pipelines were proposed using Laplacian of Gaussian operators in generating regions of interest [36] and using FP reduction methods as a postprocessing step [12]. While two-step methods proved effective, CNN based single pass methods such as modified GoogLeNet [14] and DeepMedic [13] are preferred because of its simplicity. To achieve performances on par with two-step methods, our work explores enhancements to the network from an architecture and training strategy standpoint.

In evaluating the performance of the BM detection methods, it is found that the FP improved significantly when lesions below 10mm were excluded [14], and similar observations in missed detections were noticed in [13], [37].

This proves that smaller BMs are difficult to detect, with its poor contrast, minute size and similarity to blood vessels. Because of the unmistakable parallels of size and multiplicity in lung nodules and BM, we adapt a competition winning lung nodule detector architecture as the basis for brain metastases detection [38].

Table I Overview of BM Segmentation/Detection Approaches Using CNNs

Authors	MRI Sequence	Number of patients	BM diameter (mm)	Sensitivity	Reported FP/scan
Jünger et al., 2021 [13]	Multi seq.[a]	98 (Train 66, Test 17)	N/A[d]	85.1	1.5
Zhang et al., 2020 [12]	Axial postcontrast 3D T1-weighted imaging	121 (Train 73, Test 48)	N/A	96	20
Grøvik et al., 2020 [14]	Multi seq.[b]	156 (Train 105, Test 51)	N/A	53	8.3
Xue et al., 2020 [34]	3D-T1-MPRAGE	1201, 4-fold CV	N/A[e]	100	N/A
Dikici et al., 2019 [36]	Gadolinium T1-weighted 3D	158, 5-Fold CV	Mean: 5.4 Median: 4.6	90	9.12
Charron et al., 2018 [39]	Multi seq.[c]	182 (146 Train, 18 Validation, 18 Test)	Mean: 8.1 Median: 7	93	7.8

[a] T1-/T2-weighted, T1-weighted contrast-enhanced (T1CE; gradient-echo and spin-echo sequences), and FLAIR
[b] Pre- and post-gadolinium T1-weighted 3D fast spin echo (CUBE), post-gadolinium T1-weighted 3D axial IR-prepped FSPGR (BRAVO), and 3D CUBE fluid attenuated inversion recovery (FLAIR) sequences.
[c] T1-weighted 3D MRI with Gd injection, T2-weighted 2D fluid attenuated inversion recovery MRI and T1-weighted 2D MRI sequences.
[d] Test set mean volume: 0.99 ± 4.24 cm^3 (translates to approximately 9.96mm diameter)
[e] Minimum diameter 2mm

2.3 The role of Attention and Non-local blocks in Deep Learning Architectures

2.3.1 Attention mechanisms

As GPU capacity increases, deeper and more complex architectures inspired by these models are being built. However, in the interest of efficiency and energy conservation, less complex models, which can provide similar accuracy without the additional computational overhead, are preferred. One promising concept is the Attention mechanism, which was initially used in natural language processing [40], but has found its way into medical image segmentation tasks. Attention gates (AG) are embedded into the skip connections of UNET to highlight regions of saliency by incorporating spatial information from coarser scales to drive the upsampling operation [41]. Similarly, Squeeze and Excitation (SE) layers were introduced to compute channel-wise importance of feature maps [42]. Both methods suggest spatial and channel information with higher information density are assigned a higher weightage. However, AG and SE layers rely on stochastic methods and global average pooling to calculate the spatial and channel importance. Pooling operations may extract regions with the highest intensities, which may not correlate with regions of the highest anomalies. This is particularly true for the brain region, where some normal structures of the brain can exhibit high contrast on the MRI.

The main purpose of attention blocks is to bias the network to regions which are densely informative. Attention gates rely on convolutions and downsampling from coarser scales to extract salient features [43]. Another popular and efficient approach - Squeeze and Excitation - uses global average pooling to determine channel-wise attention

[42]. Here, spatial information in each channel is compressed into a single weight, so that significant areas may be ignored. However, this may be detrimental for brain tumor segmentation tasks, since tumorous areas may be inadvertently left out, in favour of higher intensity regions. While the original SE layers focus on channel-wise attention, this was extended to include spatial and channel squeeze and excitation in [44]. The proposed Attention UNET architecture added AG to skip connections in the expansive path, neglecting to incorporate attention to the contracting path [43]. While these methods have used fully-connected and pooling layers to obtain averages of spatial or channel information, they may not reflect the exact tumorous regions.

Despite advancements in deep learning methods, techniques borrowed from classical machine learning work can be incorporated to further improve segmentation accuracy. Classical brain tumor segmentation methods mimic radiologists by utilizing the assumption of bilateral asymmetry of the brain to localize tumor regions [45][46]. One method uses fixed templates to determine asymmetry, which may not be able to account for structural variations in patients [45]. Another method requires an estimate of the mid-sagittal plane, which would introduce some uncertainty since there could be structural differences between patient scans [46]. In [47], the shortcomings described in the earlier methods are addressed by assuming a simple contralateral flip of feature maps. However, the method assumes feature maps across channels carry similar information, negating any channel importance emphasized in the SE blocks.

Our method proposes to use the inherent asymmetry in tumorous organs [48][10][23] as attention maps. The aim of incorporating this asymmetric information in network layers is to highlight boundary regions which have proved difficult to segment due to lack of contextual information [49]. However, the authors in [47] highlighted that while the flip method fared better for whole tumor structures, it did not perform well on finer structures such as enhancing tumors. To address this, we incorporate multidimensional asymmetry, through a dual scale asymmetric attention scheme [50][51] calculated on three planar views to provide additional spatial context to the attention map.

2.3.2 Non-local blocks

Deep learning architectures consist of convolutional operations which are primed to extract information from its local neighborhood. In 3D medical applications, convolutional operations may not be sufficient to capture information of objects of interest that are volumetric in nature. In this regard, Attention and Squeeze and Excitation methods highlight regions of interests through saliency maps [40], [42]. However, these methods evaluate an overall map from feature maps, which is less precise than non-local blocks as the latter model long-range dependencies at each pixel position [52]. The non-local block's approach to model a pixel's response in relation with the entire volume is advantageous in detecting BMs as presence of small lesions are observable across a few slices [52].

2.4 Methods in Semi-supervised learning

Deep neural networks require large amounts of labeled data, while expert labeling is costly and time-consuming. Conversely, data that is unlabeled is abundant and easily obtainable. To this end, semi-supervised learning methods are used to train networks with small amounts of labeled data and large portions of unlabeled data. For semi-supervised learning in medical image analysis, they fall into two broad categories, one for "self-training" and the other for regularization via graph-based methods and Support Vector Machines [53].

One of these methods is pseudo-labeling, which uses predicted values of unlabeled data for entropy regularization, enabling the network to generalize better and to reduce classification error [54]. In pseudo-labeling, the network is trained with labeled data as it usually would, and pseudo labels are used in the fine-tuning phase. In the fine-tuning phase, the network will be used on unlabeled batches to obtain pseudo labels. Subsequently, the training will be performed by evaluating the loss on the pseudo-labeled batch and labeled data (see Figure 1). The weighted pseudo-labeled loss acts as a regularization term, and it is slowly increased as the number of epochs increases.

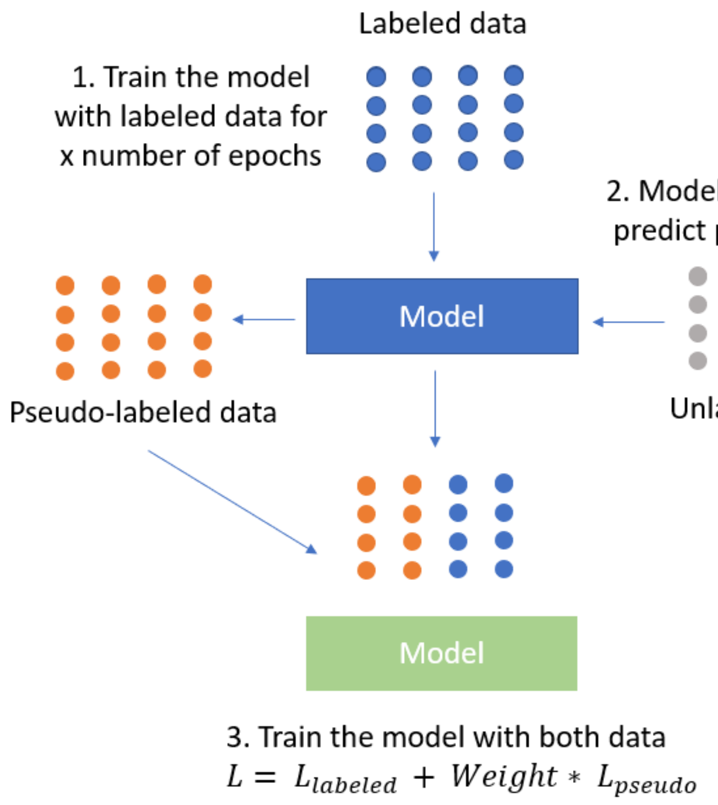


Figure 1 Flowchart of the pseudo-labeling method

Authors have incorporated a Conditional Random Field into the pseudo labeling method in MRI cardiac images with vast improvements to the Dice score [55]. Errors in a misclassified pseudo-label tend to get augmented by the network in successive iterations. By training networks on coronal, axial, and sagittal planes, these errors can be minimized [56]. A more straightforward approach of using soft pseudo labels showed that ensuring a minimum amount of labeled data in a batch and mix-up augmentation is adequate in overcoming confirmation bias [57]. Taking a different approach, these authors generate pseudo labels using nearest neighbor graph-based methods rather than relying on predictions on the network itself [58]. Decoupling the models used for pseudo-labeling and training aims at reducing the confirmation bias faced by traditional pseudo-labeling schemes.

2.4.1 Data scarcity in Medical Datasets

In deep learning applications for medical images, fully annotated data is insufficient to train the network to a good accuracy. To alleviate this, models are pre-trained from thousands of natural image datasets [59]. However, transfer learning from natural image to medical imaging tasks did not appear to improve performance [60], as natural images are inherently different from medical images. Highly specialized transfer learning where the source data and target are medical applications, seems to fare better [61].

To increase data volume in the training set, auxiliary datasets are jointly used in deep learning applications. Datasets with small domain gaps tend to perform better when trained together, and domain adaptation is key to narrow the domain gap. Another source of data comes from unlabeled medical images, which are pseudo labeled via self-training paradigms. Because self-learning is iterative in nature, achieving good accuracies require few iterations, which is costly and time consuming [62]. Ultimately, the law of diminishing returns prevails after 3-4 iterations [63]. Another unknown is the uncertainty of the pseudo labels, which has been a topic of deep learning research [64]. Few-shot object detection methods use small data volumes, but the accuracy achieved is still far from those in classic object detection [65].

2.4.1.1 Semi-Supervised Learning in Medical Datasets

The acquisition of fully-annotated medical images for supervised learning for segmentation tasks is limited and costly, compared to the availability of unlabelled images. To overcome this, transfer learning on encoders trained with out-of-domain ImageNet data was proposed for brain tumor segmentation [66]. However, as only three modalities were used in finetuning to conform with the RGB channels of a pretrained ResNet34 encoder, results did not favour a four-modality MRI dataset. Pseudo labelling was used to train labelled and unlabelled data together with a weighted unlabelled loss which increased with epochs, as the confidence of the pseudo labels improved [67]. In brain image and Covid-19 segmentation tasks, pseudo labels were augmented with MixUp with data from hidden layers in [68]. The Noisy Student method showed that adding noise during self-training improved classification tasks [69]. This study was extended to semantic segmentation, which showed that self-training helps accuracy where pre-training may not [62].

Much success has been demonstrated in self training for classification, particularly for large datasets. The application of self-training to semantic segmentation is briefly explored in [62]. However, for brain tumor segmentation, the Noisy Student method has yet to be explored. In line with the academic theme of teacher-student training, we incorporate the idea of curriculum learning to the teacher-student iterations. Curriculum learning advocates learning tasks with increasing difficulty to improve generalization of the network [70]. The difficulty of a sample could be determined by the disease severity classified by a physician [71]. However, there could be an inverse relationship between disease severity and the sample difficulty. Another easy-to-difficult strategy involves training the network in a coarse-to-fine fashion, to expose the network to large tumor regions before moving on to finer structures [49].

Here, we propose a simple yet effective way to increase the complexity of our data through the addition of noise. Our teacher model starts off with training on only labelled images from the dataset. Subsequently, we inject uncertainty by joint training with pseudo labels, and we progressively add more augmentations so that the model can learn more difficult scenarios.

2.4.1.2 Domain Adaptation in Medical Datasets

It is widely accepted that to sufficiently train the neural network, large volumes of annotated data are required to achieve good results. The original Faster-RCNN paper was benchmarked against the PASCAL VOC 2007, which contained 5000 training and 5000 test images [10]. In contrast, a Faster-RCNN used in BM detection used 73 patients for training and 48 for testing [12]. The scarcity of data is more pronounced in medicine, from concerns of ethics in data collection to getting a team of radiologists to dedicate time on top of their daily workload to process the MRI images in the dataset. Models Genesis demonstrated that a model trained on chest CT images contained domain-transferable qualities for other different domain medical datasets [61]. However, same-domain training is favored over training on different domains because of the smaller knowledge gap in learned representations.

Model performance is measured by its ability to generalize on unseen test data. Oftentimes, a model trained on a homogenous data source will not perform well when presented with data from different acquisition protocols or different hospitals [72]. This phenomenon is common in medical data where different contrast, settings, machines, and modalities could cause domain shifts. Self-training paradigms that incorporate unlabeled data and auxiliary datasets have been shown to narrow these domain gaps [61], [73]. To reduce joint-training domain conflict, dynamic convolutional networks can be employed to overcome the domain shifts, but it is computationally intensive [74].

3 Datasets and Overview of Preliminary methods

3.1 Overview

There are four datasets, which constitute all the datasets involved in this thesis. Initial work was carried out on the BraTS dataset for segmentation tasks while unlabeled data in the validation set was used on curriculum learning mechanisms. Subsequently, for the task of brain metastases detection on SUNMET, BraTS was used as a helper dataset as it was found that co-training led to performance improvements [26]. To study domain adaptation of large and small domain gaps, two more datasets, BrainMetShare and DUKE Breast Cancer dataset were used in gradual self-training and data fusion methods. Table II contains information of all 4 datasets to be discussed in this thesis.

3.1.1 Declarations

The internal ethics review board approved the use of the dataset in this study (reference number: 2019-19668-35727). Since the study was retrospective, the requirement for written informed consent was waived by the ethics committee of Sunway Medical Centre.

Table II Distribution of patients' demographics in the BraTS, DUKE and SUNMET dataset (age information has been redacted for patient confidentiality)

	BraTS	BrainMetShare	SUNMET	DUKE
Total number of patients	1251	156	121	910
Gender distribution	Male: N/A Female: N/A	Male: 51 Female: 105	Male: 56 Female: 65	Female: 910
Organ affected	Brain	Brain	Brain	Breast
Cancer Type	Primary gliomas	Brain Metastases	Brain Metastases	Breast cancer
Primary cancer	Glioblastoma multiforme:1251	Lung:99 Breast:33 Skin/melanoma:7 Genitourinary:7 Gastrointestinal:5 Miscellaneous:5	Breast:3 Pituitary:3 Lung: 12 Skin:1 Spinal cord:2 Colon:1 Cancers of Unknown Primary (CUP):99	Breast:910
Lesion sizes	≤7: 541 (58%) 8-12: 44 (5%) >12: 349 (37%)	≤3: 64 (41%) 4–10: 47 (30%) >10: 45 (29%)	≤7: 338 (3%) 8-12: 150 (51%) >12: 231 (46%)	≤7: 0 (0%) 8-12: 24 (3%) >12: 882 (97%)

*Numbers indicate the number of patients

3.1.2 SUNMET dataset

On SUNMET, post-gadolinium T1-MPRAGE sequences were interpreted by a clinical radiologist with 10 years of experience on a Carestream workstation (New York, United States). The lesions were manually marked using "Response Evaluation Criteria In Solid Tumors" (RECIST) guidelines on key slices and saved as DICOM Key Images for future reference (3). Depending on the size and shape of the metastases, every metastasis was described by one or two of RECIST diameters accompanied with a length (in mm or cm) to measure the lesion extent.

3.1.3 BraTS Dataset

Gliomas are the most common primary brain tumors in adults [75]. High-grade gliomas (HGG) are rapidly growing tumors with a median survival rate of 2 years and require immediate medical attention [76]. Conversely, low-grade gliomas (LGG) exhibit slower development and treatment is often delayed. Irrespective of type, both require extensive neuroimaging scans for diagnoses, treatment planning and for post-operative evaluations – which are mainly assessed by a clinician qualitatively [27].

However, there was no widely accepted method to evaluate the performance of brain tumor segmentation methods [77]. In 2012 and 2013, twenty teams participated in the Multimodal **Bra**in **T**umor Image **S**egmentation (BraTS) Benchmark challenge was held in conjunction with the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference. To evaluate the accuracy of the participating algorithms, a dataset consisting of MRI scans of LGG and HGG patients with ground truth and a hidden test set was prepared. This dataset was made available online and teams could submit their segmentation results for evaluation online. The challenge is now held annually and the dataset has undergone many iterations and additions [27], [77], [78].

BraTS data was manually annotated into four regions (namely edema, tumor core, enhancing core and non-enhancing core) by up to four experts using a standardized annotation protocol given in [79] using co-registered T1,T1c,T2 and FLAIR MRI contrasts.

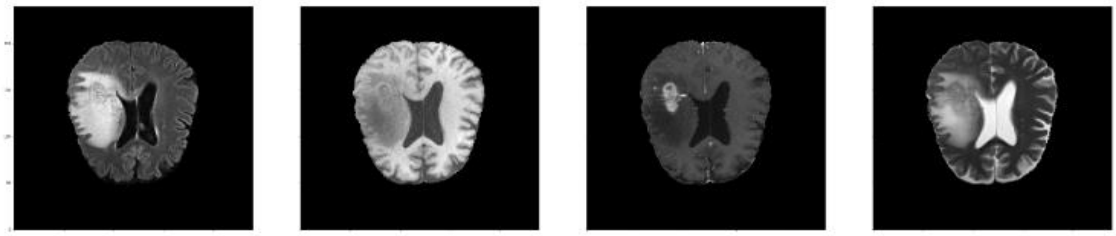


Figure 2 BraTS MRI data in four different modalities (Left to right): (1) FLAIR, (2) T1, (3) T1c (4) T2

The images in the dataset all have the following modalities as shown in

Figure 2 BraTS MRI data in four different modalities (Left to right): (1) FLAIR, (2) T1, (3) T1c (4) T2 [27]:

1) T1: T1-weighted, native image, sagittal or axial 2D acquisitions, with 1–6 mm slice thickness.
2) T1c: T1-weighted, contrast-enhanced (Gadolinium) image, with 3D acquisition and 1 mm isotropic voxel size for most patients.
3) T2: T2-weighted image, axial 2D acquisition, with 2–6mm slice thickness.
4) FLAIR: T2-weighted FLAIR image, axial, coronal, or sagittal 2D acquisitions, 2–6 mm slice thickness.

All data are co-registered to the T1c MRI, skull stripped and isotropically normalized to $1mm^3$. Four multiclass labels are provided for the training of the segmentation algorithms (0 – background, 1 – Necrotic core, 2- Edema, 3- Non-Enhancing 4 – Enhancing Tumor)[1].

However, during evaluation, these labels are categorized into tumor regions which are more relevant to clinical applications, namely [27][77]:

1) the "whole" tumor (WT) region (union of all labels i.e., 1, 2, 4)
2) the tumor "core" (TC) region (union of 1 and 4)
3) and the "active" tumor (ET) region (only containing the "enhancing core" structures that are unique to high-grade cases).

Based on the evaluation of the manual segmentations, it is found that the WT region was easily identifiable in both HGG and LGG cases, while the TC regions in LGG and ET regions in HGG were more difficult to determine. As the BraTS dataset is constantly updated, as of time of writing, this thesis will cover the use of BraTS 2018, 2020 and 2021.

[1] Label 3 has been eliminated beginning from BraTS 2017 as there was potential it could have been mislabeled

3.1.3.1 Class Imbalance in BraTS dataset

In medical images, the amount of healthy tissue is more substantial than lesions, leading to prediction outcomes that favor the majority class [80]. Indiscriminate application of machine learning methods without consideration of class imbalance leads to unfeasible results [81].

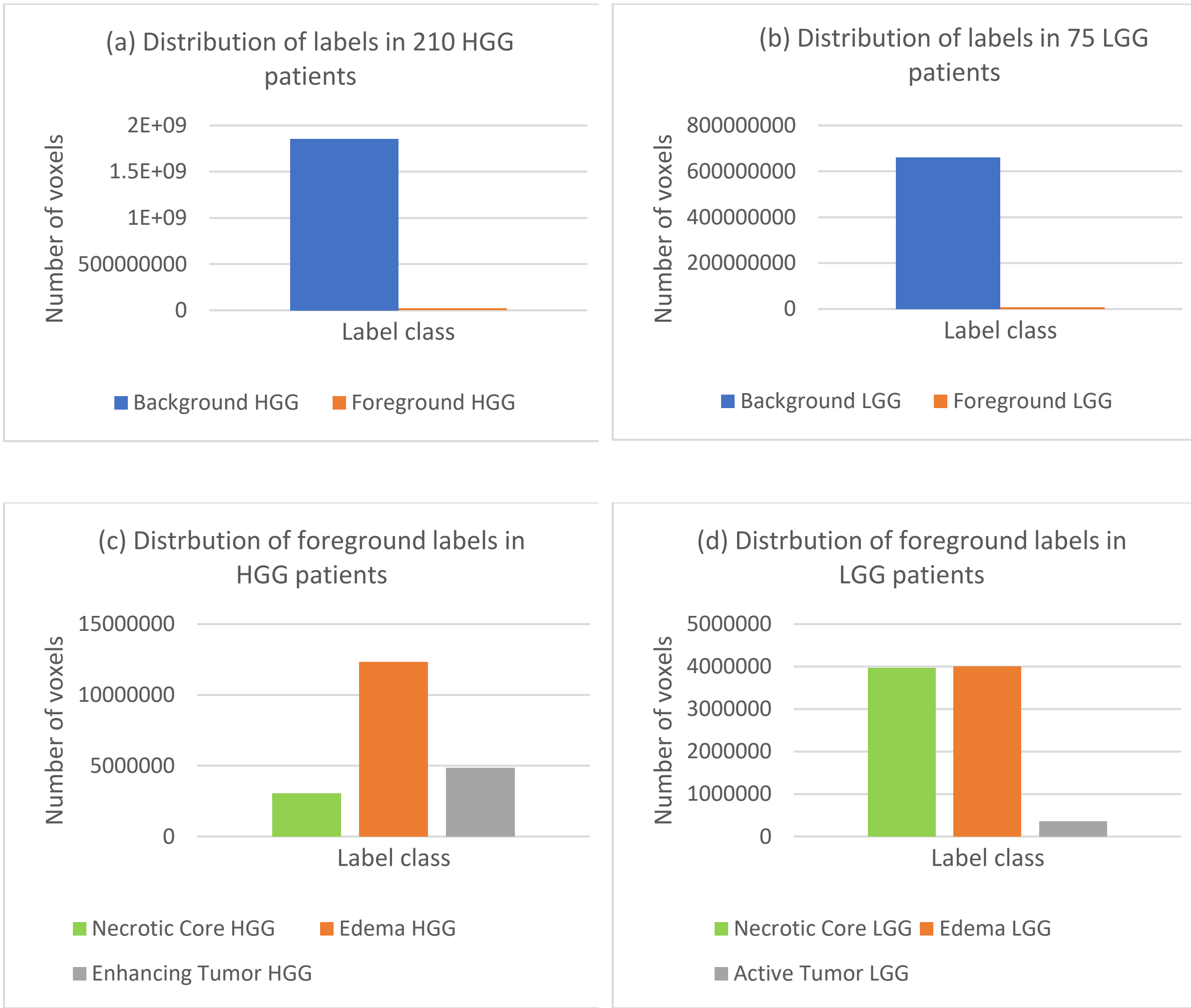


Figure 3 Number of labels per class in HGG and LGG data from the BraTS 2018 datasets

BraTS data suffers from a similar fate where there are more background and healthy tissue than lesions. As can be seen in Figure 3, there is at least a thousand times more background than the other classes combined. For a deep learning algorithm learning to generalize, it would seem obvious to predict all pixels as background to minimize the loss function. Once again, subplots (c) and (d) show an apparent imbalance between foreground classes (labels 1,2, and 4).

To counter class imbalance approaches such as modifying learning methods, employing cost-sensitive learning, and resampling strategies are proposed [82]. The SMOTE algorithm is a popular resampling strategy that suppresses majority classes and oversamples the minority class through data augmentations for a balanced dataset [83]. Using the SMOTE technique achieved higher Dice Coefficients for the core and enhancing tumor regions in the BraTS dataset, corresponding to better performance in the minority class [84]. However, this technique may inadvertently remove pertinent information or introduce redundancy into the training sets.

In medical image segmentation, a standard approach to counter class imbalance is to use patch-based networks [85]. It is paramount that effective sampling strategies are used with patch-based networks to obtain equal distribution between background and foreground regions. DeepMedic's patch strategy selects segments with a center pixel that has a 50% probability of being foreground or background [6]. For the segmentation of white matter lesions, the lesion centered strategy dictates that patches contain white matter lesions anywhere within the area [85]. A similar procedure is used in ischemic stroke lesion segmentation [86].

Patch labeling strategies are often combined with weighted loss functions, such as the Generalised Dice Loss [87] or Generalised Wasserstein Dice Score [88] or multi-step training [11] to counter the class imbalance. In this multi-step training, equiprobable patches containing all labels (labels 0,1,2, and 4 in BraTS) are selected in the first phase of training. The next step uses data representative of the imbalanced dataset. In conditional Generative Adversarial Networks (GAN) for brain tumors, weighted loss functions are combined with the adversarial loss to counter the class imbalance [30].

3.1.4 BrainMetShare Dataset

To date, only a handful of work using deep learning on BM can be found in [12]–[14], [36]. The speed of progress in primary brain cancer detection methods compared to BM detection can be attributed to the widely available BraTS dataset and annual challenge [78], [79], [89]. Recently, a publicly available dataset on brain metastases was released by Stanford University, which contains 156 patients (105 for the training and 51 for the validation set) with at least 1 cerebral metastasis [14].

BrainMetShare dataset was annotated by two neuroradiologists with 8- and 2-years' experience by manually delineating regions of interest on each slice of post-Gd 3D T1 weighted IR-FSPGR sequence with guidance from 3D FLAIR and post-Gd T1-weighted spin echo on OsiriXMD software package (v. 8.0, Geneva, Switzerland) (1)

The dataset contains four sequences of volumetric MRI images; pre- and postgadolinium T1-weighted 3D fast spin echo (CUBE), post- gadolinium T1-weighted 3D axial IR-prepped FSPGR (BRAVO), and 3D CUBE fluid attenuated inversion recovery (FLAIR) (See Table for demographic data). In contrast with the BraTS dataset, the ground truth uses a single label to annotate the entire tumorous region.

3.1.5 DUKE Breast Cancer Dataset

For the Duke Breast Cancer dataset, the ground truth was given by (row_start, row_end, column_start, column_end, slice_start, slice_end) and additional preprocessing was performed to convert it to centroid information. In terms of the MRI sequences, since the scans came from various centers, we randomly selected the MRI modalities to introduce some variance in the training data. Also, it is important to note that some of the image sequences were ordered differently based on the ImagePositionPatient DICOM tag. The images were normalized so that the start_slice is always at the lowest height, and this was verified manually by visual inspection of all the ground truths (see **Figure 4**)

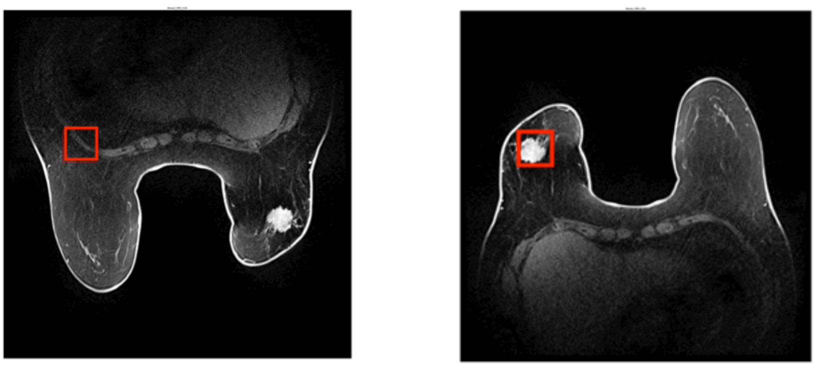


Figure 4(a) shows the ground truth with start_slice at the highest z slice (b) after realignment, notice the ground truth is now situated within the lesion

3.2 Interobserver Variability in ground-truth labeling of datasets

It is a generally accepted notion that interobserver variations occur between experts. Firstly, this presents challenges in surgical planning and diagnosis of brain tumors. In deep learning, misrepresented regions of interest introduce noise into the ground truth, destabilizing the deep learning model since backpropagation is based on the ground truth of the training set to arrive at the best possible solution. Another pitfall would be in the performance measurements obtained on the validation set.

To alleviate the impact of interobserver variations in brain metastases assessment, consensus or expert panel, multiple reading, the use of standardized reporting guidelines, interdisciplinary conferences or tumor boards involving multiple specialists, and specialized training programs are employed.

The SUNMED dataset has been obtained from standard clinical practice from a clinical oncologist with 10 years' experience. Due to limited resources, further consensus from other experts or multiple readings could not be obtained for the SUNMED dataset. For this reason, all analysis conducted on the SUNMED dataset are supported by results from the BrainMetShare dataset, which was annotated and cross checked by 2 experienced neuroradiologists, with 8 and 2 years of experience, respectively. We have found that in general, the improvements and training paradigms which we have introduced show similar trends in performance gains/reductions across all datasets, which reinforces the applicability of our methods.

3.3 Overview of the Preliminary Methods

Figure 5 is an overview of the proposed deep learning architectures and datasets associated with training them. Arrows indicate that some proposed concepts are refined in the next iteration of the system to improve the performance of the model. Note that each architecture exists as a standalone model with brain lesion detection or segmentation capabilities.

Chapter 4 begins with the segmentation of brain tumors using a CASPIANET++ which employs a multiview approach and evaluating brain symmetry in determining regions of saliency in brain MRI – which achieves performance comparable with competition winning architectures. A noisy student curriculum learning is introduced to harness the potential of unlabeled BraTS validation data.

In Chapter 5, a lung nodule detector was repurposed into a BM detector because of the physical parallels between brain metastases and lung nodules. Non-local blocks were added to provide far-range information as metastatic lesions affect distant slices that is beyond the capacity of local convolutional operations. From 3 datasets, a gradual self-training paradigm was proposed, and it was demonstrated that the method was able to generalize on both BM and primary glioma datasets.

In Chapter 6, it is postulated that while non-local blocks had improved sensitivity, the FP rate could be further reduced, by implementing a near-range non-local block based on the nature of the small metastatic lesions. Instead of an iterative training approach of pseudo labels, we use the ground truth from different organs and different pathologies to improve the performance of the BM detection.

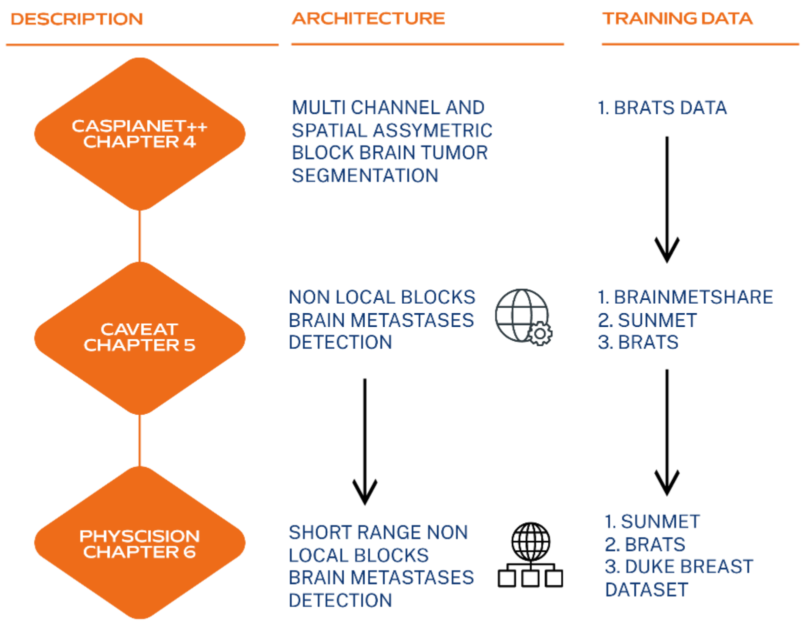


Figure 5 Overview of the deep learning ecosystems presented in this thesis and the training data used in those particular studies, arrows indicate continuity of an idea carried to the next approach

3.4 General statement on selecting the best training epochs.

All experiments in Chapters 4, 5 and 6 are conducted with fixed training epochs, using an adaptive learning rate policy. The models are check-pointed every 10 epochs, to be evaluated on the validation dataset. The final model is selected based on the best results obtained on the validation dataset. For the BraTS dataset, this requires multiple online submissions to determine, as ground truth is unavailable on the validation set – to maximize the use of all ground truths from the training set. For SUNMED and BrainMetShare, the validation set is a 5-fold cross validation dataset randomly generated via sklearn library. We selected the validation set that gave the median performance on the base model.

These steps are taken to ensure that the most optimum solution is found, as there are variations in each architecture, training data, and possibility of overtraining, which may affect the number of epochs it takes to reach the best solution.

4 CASPIANET++: A Multidimensional Channel-Spatial Asymmetric Attention Network with Noisy Student Curriculum Learning Paradigm for Brain Tumor Segmentation

4.1 Key points

- Morphological information in brain structures can be developed as attention maps
- Multidimensional channel spatial approaches provide additional information to deep learning architectures
- Progressively incorporating noise in training the dataset ensures that the model learns in a curriculum-learning approach

4.2 Introduction

Convolutional neural networks (CNNs) have been used quite successfully for semantic segmentation of brain tumors. However, current CNNs and attention mechanisms are stochastic in nature and neglect the morphological indicators used by radiologists to manually annotate regions of interest. In this chapter, we introduce a channel and spatial wise asymmetric attention (CASPIAN) by leveraging the inherent structure of tumors to detect regions of saliency. To demonstrate the efficacy of our proposed layer, we integrate this into a well-established convolutional neural network (CNN) architecture to achieve higher Dice scores, with less GPU resources. Also, we investigate the inclusion of auxiliary multiscale and multiplanar attention branches to increase the spatial context crucial in semantic segmentation tasks.

This chapter describes:

1. A novel Channel Spatial Asymmetric Attention (**CASPIAN**) Layer based on a-priori asymmetry information, where highly dissimilar regions are given higher importance and propagated through the network.
2. A multidimensional approach integrating multiscale and multiplanar (i.e., axial, coronal and sagittal) information into the CASPIAN layer, resulting in the **CASPIAN++** (CASPIAN + Multiscale + Multiplanar) layer, to increase spatial information within the network
3. An improved deep neural network architecture, **CASPIANET++** which integrates multiscale and multiplanar spatial attention with channel attention for more accurate segmentation
4. A curriculum learning paradigm called **Noisy Student Curriculum learning** for brain tumor segmentation tasks

4.3 Methods

Spatial and Channel Asymmetry share the same input I, which is the output of any arbitrary encoder or decoder of the CNN architecture. All the inputs in this section are feature maps with dimensions $\mathrm{HxWxDxC}$, where $\mathrm{H, W, D, C}$ correspond to the height, width, depth and number of channels, respectively.

4.3.1 Spatial Asymmetric Attention Mask

Let feature maps obtained from a layer be U, consisting of feature vectors $\{u_{ijk}^1, u_{ijk}^2, \dots \dots u_{ijk}^C\}$ for a single point i, j, k along the entire channel axis from 1 to C, where $i = 1$ to $H, j = 1$ to $W, k = 1$ to D. Similarly, $\widehat{U}$ is the horizontal-flip of U, with corresponding feature vectors $\{\hat{u}_{ijk}^1, \hat{u}_{ijk}^2, \dots \dots \hat{u}_{ijk}^C\}$.

$\alpha_{sp}(i, j, k)$, is a similarity measure, which quantifies how similar a particular pixel at location i, j, k is relative to its horizontally flipped counterpart. It is defined as:

4.3.1.1.1

4.3.1.1.2 $$\alpha_{sp}(i,j,k) = \frac{\sum_{l=1}^{C} u_{ijk}^{l} \hat{u}_{ijk}^{l}}{\sqrt{\sum_{l=1}^{C} {u_{ijk}^{l}}^2} \sqrt{\sum_{l=1}^{C} {\hat{u}_{ijk}^{l}}^2}}$$ 4.3.1.1.3 (1)

$\alpha_{sp}(i, j, k) \in \mathbb{R} \mid [-1,1]$. Values closer to -1 and 1 indicate high dissimilarity and high similarity, respectively.

Values of α_{sp} in the range [-1,1] are normalized to [0,1] to obtain θ_{sp}, the Spatial Asymmetric Attention Mask (SAAM)

$$\theta_{sp} = \frac{1 - \alpha_{sp}}{2} \tag{2}$$

Dissimilar pixels indicate that the symmetry rule in the left and right regions have been violated and there is a high chance of tumor in that pixel. After normalization, regions of high dissimilarity have higher values and can be represented as a weighted mask, to give attention/importance to the tumorous regions.

4.3.2 Channel Asymmetric Attention Mask

In [47], every channel was assumed to contain similar spatial information and carries equal importance. However, different tumor regions tend to exhibit better contrast in certain MRI modalities. A similar protocol is followed by experts in annotating the ground truth for the BraTS challenge [27]. Therefore, to account for the multimodal nature of the images, we extend the spatial attention method in [19] and propose to weight each channel by a scalar value, termed channel-wise asymmetric attention.

Let feature maps obtained from a layer be M, consisting of feature vectors $\{m^p_{1,1,1}, m^p_{1,1,2}, \dots \dots m^p_{H,W,D}\}$ for every point in the feature map at a single channel axis p, where $p = 1$ to C. Similarly, $\widehat{M}$ is the horizontal-flip of M, with corresponding feature vectors $\{\widehat{m}^p_{1,1,1}, \widehat{m}^p_{1,1,2}, \dots \dots \widehat{m}^p_{H,W,D}\}$.

$\alpha_{ch}(p)$, is a similarity measure, which quantifies how similar a particular channel p is relative to its horizontally flipped counterpart. It is defined as:

$$\alpha_{ch}(p) = \frac{\sum_{ijk} m^p_{ijk}\, \widehat{m}^p_{ijk}}{\sqrt{\sum_{ijk} {m^p_{ijk}}^2} \sqrt{\sum_{ijk} {\widehat{m}^p_{ijk}}^2}} \tag{3}$$

$\alpha_{ch}(p) \in \mathbb{R} \mid [-1,1]$. Values closer to -1 and 1 indicate high dissimilarity and high similarity, respectively.

Values of α_{ch} in the range [-1,1] are normalized to [0,1] to obtain θ_{ch}, the Channel Asymmetric Attention Mask (CAAM) as follows:

$$\theta_{ch} = \frac{1 - \alpha_{ch}}{2} \tag{4}$$

The resulting θ_{ch} is a scalar value denoting how dissimilar the feature map for each channel is, indicating which channel may hold more salient information.

4.3.3 Channel and Spatial Asymmetric Attention (CASPIAN) Layer

In the literature, brain tumor segmentation methods rely solely on the bilateral asymmetry on the mid-sagittal plane regions to pinpoint the locality of tumors. This necessitates that the mid-sagittal plane be as accurate as possible. However, in our method we apply a geometric flip to the input feature maps to calculate its asymmetry mask. The geometric flip is merely a simplification to determine the bilateral asymmetry in lieu of calculating the mid-sagittal plane. The oversimplification in the left-right flip may however result in some unwanted regions of asymmetry. To overcome this, in the final step we multiply the asymmetry mask with the feature maps, to ensure that the high intensity from the tumor regions amplify the regions of asymmetry while the low intensity non-tumor regions attenuate the unwanted regions.

The similarity masks from equations (2) and (4) are scaled linearly from 0 to 1. The resulting attention mask alone, when multiplied by its input, was insufficient to amplify the salient regions. Therefore an excitation block (depicted in Figure 6), which consists of a fully connected layer, FC, a non-linear ReLU block, R, followed by another fully connected layer, FC, and sigmoid layer, σ, is first applied to the mask. While the original excitation block [42] had proposed a bottleneck layer for computational efficiency, we did not utilize a bottleneck layer to preserve the structural integrity of the asymmetric attention mask. The resulting asymmetric attention mask with excitation for spatial, $\mathrm{S}_{\text{spatial attn}}$ and channel, $\mathrm{S}_{\text{channel attn}}$ are obtained as follows:

$$\mathrm{S}_{\text{spatial attn}} = \sigma(\mathrm{FC}_2(\mathrm{R}(\mathrm{FC}_1(\theta_{\mathrm{sp}})) \qquad (5)$$

$$\mathrm{S}_{\text{channel attn}} = \sigma(\mathrm{FC}_2(\mathrm{R}(\mathrm{FC}_1(\theta_{\mathrm{ch}})) \qquad (6)$$

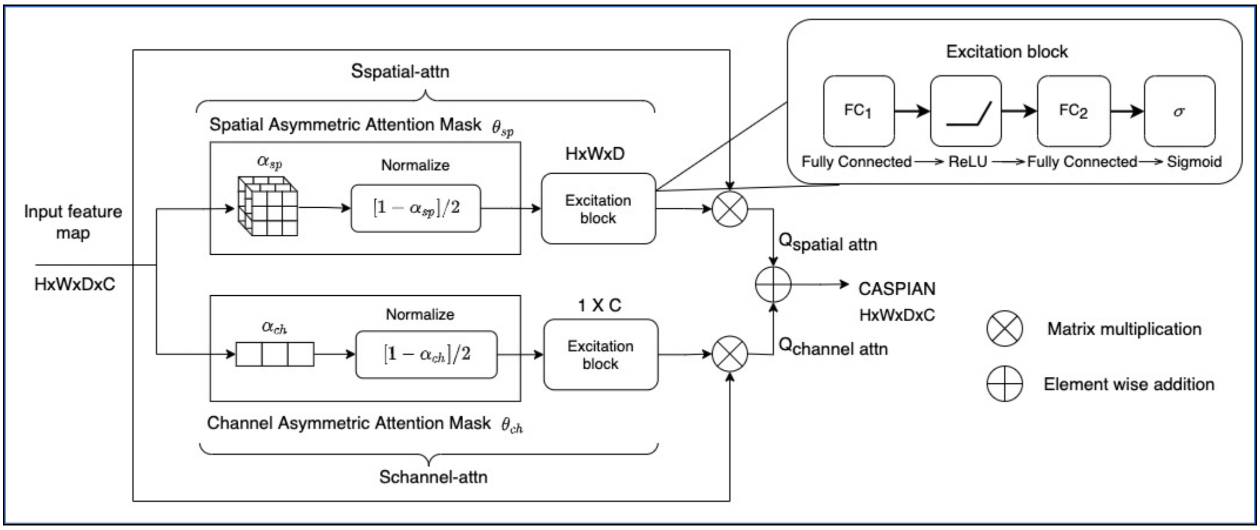


Figure 6: The proposed channel and spatial asymmetric attention with excitation block (CASPIAN), where the input feature map is the output of any arbitrary encoder or decoder of the CNN architecture

The spatial attention map $\mathrm{Q}_{\text{spatial attn}}$ and the channel attention map $\mathrm{Q}_{\text{channel attn}}$ are obtained from element-wise multiplication of the input I :

$$\mathrm{Q}_{\text{spatial attn}} = \mathrm{S}_{\text{spatial attn}} \cdot \mathrm{I} \qquad (7)$$

$$\mathrm{Q}_{\text{channel attn}} = \mathrm{S}_{\text{channel attn}} \cdot \mathrm{I} \qquad (8)$$

The $\mathrm{CASPIAN}$ block places equal importance to both channel and spatial attention, hence both maps are aggregated through element wise addition (as depicted in the final addition block in Figure 6):

$$\mathrm{CASPIAN} = \mathrm{Q}_{\text{spatial attn}} + \mathrm{Q}_{\text{channel attn}} \qquad (9)$$

4.3.4 Multiplanar and Multiscale approach in CASPIAN layers

Tumors are non-homogenous structures which have dissimilar boundaries when viewed from different planes (coronal, axial and sagittal) of the MRI data. This suggests that a similarity mask evaluated on a single view may not be adequate in capturing information for the entire volume of the tumor. We therefore introduce a Multiplanar Spatial Asymmetric Attention block as depicted in Figure 7.

From equation (5), we compute $\mathrm{S}_{\text{spatial attn,plane}}$, the spatial asymmetry mask with excitation between the feature map and its geometrical flip along a plane. $\mathrm{S}_{\text{spatial attn,plane}}$ is evaluated for sagittal (plane=1), coronal (plane=2), and axial (plane=3) planes.

$Q_{\text{multiplanar}}$, the Multiplanar Spatial Asymmetric Attention, is the sum of the element-wise multiplication of the $S_{\text{spatial attn,plane}}$ with the input image I :

$$Q_{\text{multiplanar}} = \sum_{\text{plane}=1}^{3} S_{\text{spatial attn,plane}} \cdot I \qquad (10)$$

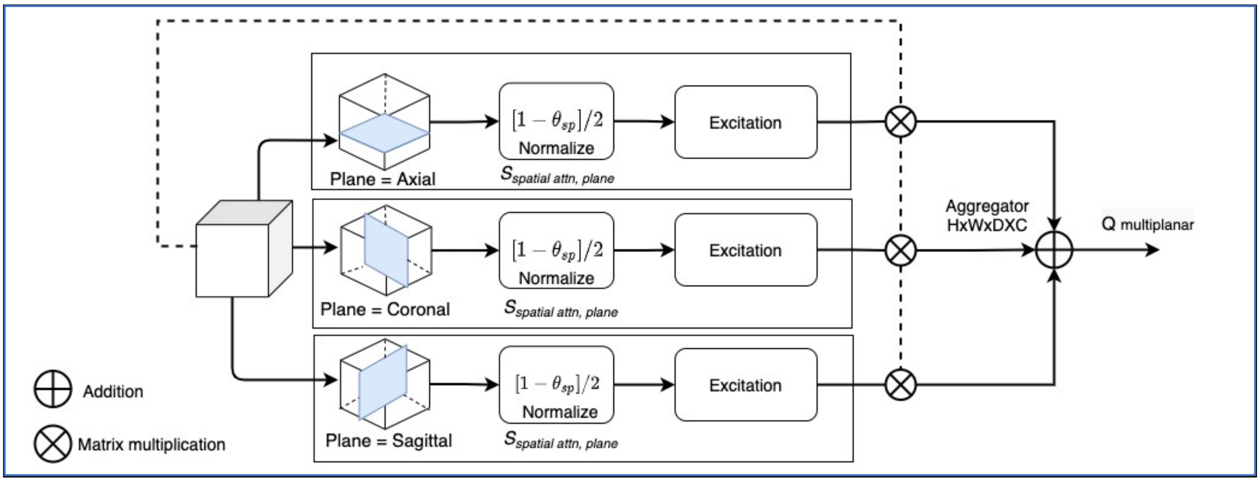


Figure 7: Multiplanar (axial, sagittal and coronal) Spatial Asymmetric Attention block

In delineating tumorous regions, the fusion of higher-level feature maps provides precise location and shape information while lower-level maps give structural context, such as texture and edges which are equally crucial in delineating tumorous regions. We therefore propose Multiscale Spatial Asymmetric Attention, $Q_{\text{multiscale}}$ to capture features across finer and coarser scales to delineate the tumors more accurately.

Consider a higher-level scale (I_{S0}), which typically is an input of the previous encoder block, and input at common scale (I_{S1}) in Figure 8. Here we introduce a gating mechanism to maximise the attention masks on both spaces. Since the most relevant information will be contained in I_{S1}, the spatial asymmetric mask $S_{\text{spatial attn,s1}}$ is given precedence.

To ensure precedence, $S_{\text{spatial attn,s1}}$ is multiplied with its image I_{S1}. To incorporate the information from I_{S0} the signal of the mask from the common scale is inverted $(1 - S_{\text{spatial attn,s1}})$ to ensure any salient regions which are missed on the common scale can be captured by the higher scale. A maxpooling operation $P(I_{S0})$ is applied to I_{S0}, so that I_{S0} is spatially equal to I_{S1}.

The Multiscale Spatial Asymmetric Attention $Q_{\text{multiscale}}$ is given by:

$$Q_{\text{multiscale}} = S_{\text{spatial attn,s1}} \cdot I_{S1} + \left(1 - S_{\text{spatial attn,s1}}\right) \cdot S_{\text{spatial attn},P(I_{S0})} \cdot I_{S1} \qquad (11)$$

In the final block incorporating multiscale and multiplanar information into the CASPIAN layer we use an additive aggregator (see Figure 9), as our empirical studies show that the max-out layer does not have a positive effect on the validation results.

The resulting multiplanar and multiscale CASPIAN layer, CASPIAN++ is defined as follows:

$$\text{CASPIAN}++ = Q_{\text{channel attn}} + Q_{\text{multiscale}} + Q_{\text{multiplanar}} \qquad (12)$$

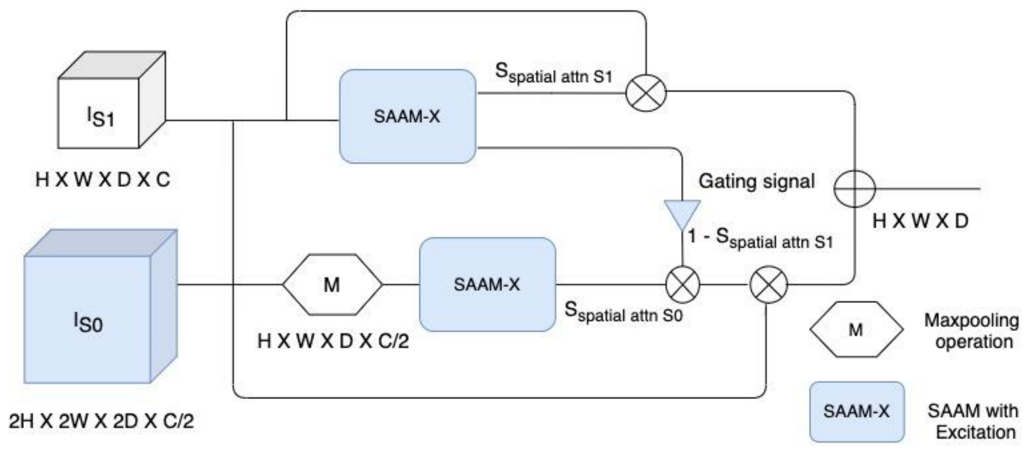


Figure 8: The proposed multiscale approach with gating mechanism

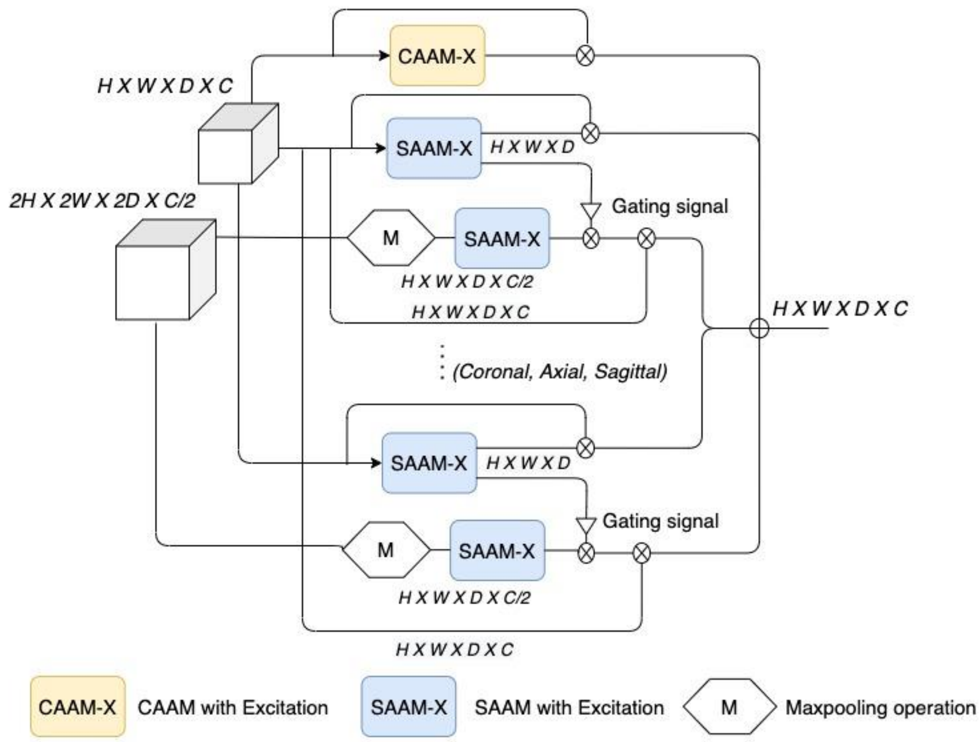


Figure 9: Combining channel asymmetric attention into the multiscale and multiplane spatial asymmetric attention block

4.3.5 CASPIANET++

The proposed CASPIAN++ layer can be integrated into any network to provide spatial and channel attention for semantic segmentation tasks. The network we chose is a lightweight variant of UNET with grouped convolutions as the base network [90]. Our architecture places CASPIAN/CASPIAN++ blocks at every path, except at the bottleneck (see Figure 10). This in effect recalculates asymmetry information for every feature map in the network, as spatial and contextual information may vary from layer to layer. A combination of CASPIAN and CASPIAN++ outputs are also added to skipped connections to filter irrelevant spatial information from earlier encoder layers for more accurate upsampling. No attention was added to the bottleneck, as spatial information has been substantially reduced from 128x128x128 to 8x8x8. Evaluating the spatial asymmetry on a layer with the least spatial information in the network would not be beneficial for segmentation tasks.

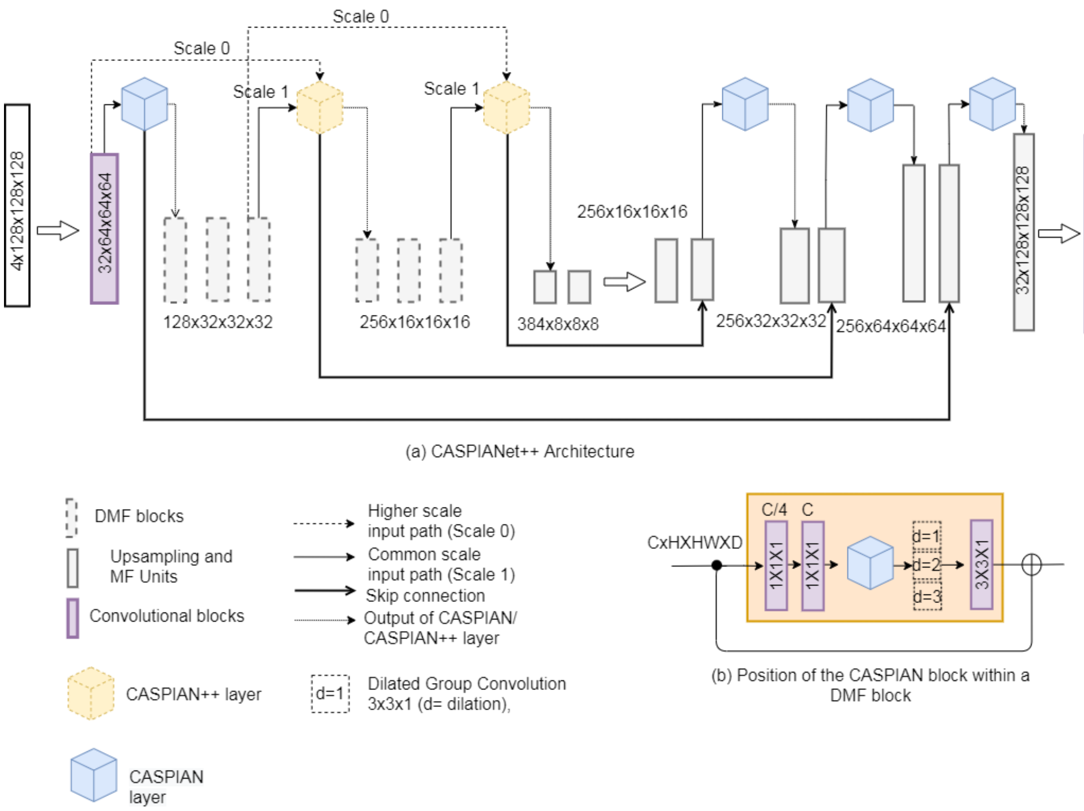


Figure 10: The placement of CASPIAN and CASPIAN++ layers in the proposed CASPIANET++. (a) CASPIAN++ layers are added to the 2nd and 3rd encoder blocks. Notice that there are no CASPIAN layers at the bottleneck. (b) depicts the addition of the CASPIAN block within the DMF (Dilated Multi-Fiber) block

CASPIAN++ was designed to include much more spatial information than the CASPIAN layer. Ideally, we would use CASPIAN++ layers in every block in the architecture. However, the multiscale approach on the first encoder block was memory intensive as it required another copy of the input image (128x128x128x4) to be stored in memory during training. Therefore, CASPIAN++ layers were added to certain layers only as a trade-off between hardware limitations and accuracy.

For maximum advantage, we incorporated CASPIAN++ layers into the second and third encoder blocks. These blocks were selected because they have the biggest receptive fields and would benefit most from the additional spatial information provided by CASPIAN++. Also, the increased spatial fields compensate for the conversion loss from spatial to contextual information. Within these blocks, CASPIAN layers are added within the DMF units to highlight salient features for the dilated convolutions (see Figure 10(b)). A similar scheme placed an attention layer after each dilated convolution [91]. However, for efficiency, we use one CASPIAN block before the dilations, to provide saliency to the dilated blocks. Hereon, the encoder-decoder implementation of CASPIAN/CASPIAN++ is known as CASPIANET++.

4.3.6 Region-based Loss Functions

The loss functions for training the network uses a region-based approach. Instead of optimizing for individual label classes (0, 1, 2, 4), the network treats the 3 regions r as three separate binary problems. The three tumor regions are defined as a combination of labels as follows: whole tumor (WT) (labels 1,2 and 4), tumor core (TC) (labels 1 and 4) and enhancing tumor (ET) (label 4).

Consider the output of a region r, where $\mathrm{g_r}$ represents the ground truth for the region and $\mathrm{p_r}$ is the sigmoid of the predicted output of the region. The Binary cross entropy $\mathcal{L}_{\mathrm{bce}}$and Dice loss $\mathcal{L}_{\mathrm{dice}}$ are defined as follows:

$$\mathcal{L}_{\mathrm{bce}}(\mathrm{g_r}, \mathrm{p_r}) = -[\mathrm{g_r}\log(\mathrm{p_r}) + (1-\mathrm{g_r})\log(1-\mathrm{p_r})] \tag{13}$$

$$\mathcal{L}_{\mathrm{dice}}(\mathrm{g_r}, \mathrm{p_r}) = 1 - \frac{2 * \mathrm{g_r} * \mathrm{p_r}}{\mathrm{g_r} + \mathrm{p_r}} \tag{14}$$

The training objective for the backpropagation step is to minimize the loss given by $\mathcal{L}_{\mathrm{total}}$

$$\mathcal{L}_{\mathrm{total}} = \sum_{\mathrm{r}=1}^{3} \mathcal{L}_{\mathrm{bce}}(\mathrm{g_r}, \mathrm{p_r}) + \mathcal{L}_{\mathrm{dice}}(\mathrm{g_r}, \mathrm{p_r}) \tag{15}$$

where $\mathrm{r} = 1,2,3$ for the respective regions of WT, TC and ET.

Assume that $\mathrm{z^l}$ is the input of the $\mathrm{l^{th}}$ layer. The backward derivative error for the backpropagation step with respect to the last layer is given by :

$$\frac{d\mathcal{L}_{\mathrm{total}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{z^l}} = \frac{d\mathcal{L}_{\mathrm{total}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{p_r}} \cdot \frac{d\mathrm{p_r}}{d\mathrm{z^l}} \tag{16}$$

$$\frac{d\mathcal{L}_{\mathrm{bce}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{p_r}} = \frac{-\mathrm{g_r}}{\mathrm{p_r}} + \frac{1-\mathrm{g_r}}{1-\mathrm{p_r}} \tag{17}$$

$$\frac{d\mathcal{L}_{\mathrm{dice}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{p_r}} = \frac{-2\mathrm{g_r}^2}{(\mathrm{g_r} + \mathrm{p_r})^2} \tag{18}$$

$$\frac{d\mathcal{L}_{\mathrm{total}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{p_r}} = \sum_{\mathrm{r}=1}^{3} \frac{-\mathrm{g_r}}{\mathrm{p_r}} + \frac{1-\mathrm{g_r}}{1-\mathrm{p_r}} + \frac{-2\mathrm{g_r}^2}{(\mathrm{g_r} + \mathrm{p_r})^2} \tag{19}$$

The $\mathrm{p_r}$ is defined as the sigmoid of the predicted output of the last layer :

$$\mathrm{p_r} = \frac{1}{1 + e^{-\mathrm{z^l}}} \tag{20}$$

$$\frac{d\mathrm{p_r}}{d\mathrm{z_j^l}} = \frac{1}{1 + e^{-\mathrm{z_j^l}}} \cdot \left(1 - \frac{1}{1 + e^{-\mathrm{z^l}}}\right) \tag{21}$$

$$\frac{d\mathcal{L}_{\mathrm{total}}}{d\mathrm{z^l}} = \sum_{\mathrm{r}=1}^{3} \frac{-\mathrm{g_r}}{\mathrm{p_r}} + \frac{1-\mathrm{g_r}}{1-\mathrm{p_r}} + \frac{-2\mathrm{g_r}^2}{(\mathrm{g_r} + \mathrm{p_r})^2} \cdot \frac{1}{1 + e^{-\mathrm{z^l}}} \cdot \left(1 - \frac{1}{1 + e^{-\mathrm{z^l}}}\right) \tag{22}$$

The gradient of the loss for the backpropagation step with respect to $\mathrm{z^0}$ at input layer 0^{th} is derived from the chain rule for all the layers of CASPIANET++ :

$$\frac{d\mathcal{L}_{\mathrm{total}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{z^0}} = \frac{d\mathcal{L}_{\mathrm{total}}(\mathrm{g_r}, \mathrm{p_r})}{d\mathrm{p_r}} \cdot \frac{d\mathrm{p_r}}{d\mathrm{z^l}} \cdot \frac{d\mathrm{z^l}}{d\mathrm{z^{l-1}}} \cdot \frac{d\mathrm{z^{l-1}}}{d\mathrm{z^{l-2}}} \cdot\cdot \frac{d\mathrm{z^1}}{d\mathrm{z^0}} \tag{23}$$

4.3.7 Noisy Student Curriculum Learning

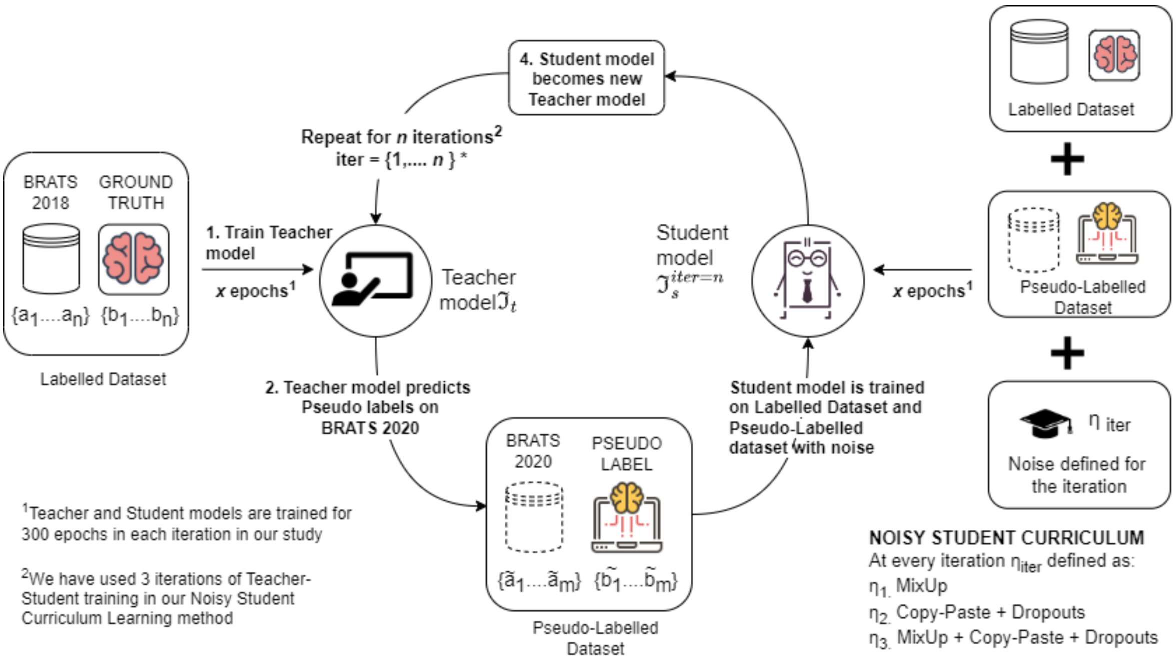


Figure 11 The Noisy Student Curriculum Learning workflow

Consider a set of images $\{(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots\ldots (a_n, b_n)\}$ where (a_n, b_n) represents the image and labels of the n^{th} patient in the dataset. The unlabeled set is denoted by $\{\tilde{a}_1, \tilde{a}_2 \ldots \ldots \tilde{a}_m\}$ where $\tilde{a}_m$ is the m^{th} unlabeled image in the dataset.

Using the curriculum learning approach, noise is successively added to the student model at every iteration so that it learns incrementally through an easy-to-hard approach; in this way, the student learns better as its parameters are first optimized on easy tasks followed by incrementally difficult tasks (see Noisy Student Curriculum method in Step 3.1 of Section 4.3.7.1). Our approach differentiates itself from the classical Noisy Student method, where the combination of all noising methods are used in every iteration from the onset of training the student model [69].

In our experiments, a student model is said to complete one iteration when trained with the labelled and pseudo labelled dataset for 300 epochs. All training data and pseudo labelled data are subjected to augmentations by centroid cropping, random rotations, intensity changes, flips and elastic deformations. This is to maintain consistency in training the teacher and student models. At the end of one iteration (step 4 of Figure 11), the student is said to “become” the teacher, as the student model is used to create pseudo labels for the next iteration. In our work, this cycle is repeated three times, after which the law of diminishing returns starts to take effect.

In our method, two data augmentation methods were utilized to noise the pseudo labels, namely MixUp [92] and Copy-Paste [93]. Copy-Paste method was chosen as it demonstrates additive effects when used with other augmentation methods [93]. To increase difficulty in the pseudo labels, we experiment with the additive nature of Copy-Paste with MixUp.

4.3.7.1 Pseudocode of our proposed Noisy Student Curriculum Learning method (depicted in Figure 11) proceeds as follows:

1. Train a teacher model $\mathfrak{I}_t$ on the labeled dataset $\{(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots\ldots (a_n, b_n)\}$ to minimize the loss function $\mathcal{L}_{total}$ in equation (15).

2. The teacher model is used to infer pseudo labels by $\{\tilde{b}_1, \tilde{b}_2 \dots\dots \tilde{b}_m\}$ for the unlabeled dataset $\{\tilde{a}_1, \tilde{a}_2 \dots\dots \tilde{a}_m\}$ as follows:

$$\tilde{b}_i = f(\tilde{a}_i, \mathfrak{I}_t), \forall\, i = 1 \dots\dots m \qquad (24)$$

3. The Noisy Student Curriculum training starts here, a single iteration $\mathbf{iter}$ is defined as a student model $\mathfrak{I}_s^{\mathbf{iter}}$ trained with labeled and pseudo labels for 300 epochs.

   3.1. Train a new student model $\mathfrak{I}_s^{\mathbf{iter}}$ with noise $\eta_{\mathbf{iter|curriculum}}$ to minimize the loss function $\mathcal{L}_{\mathbf{total}}$ on the labeled dataset $\{(a_1, b_1), \dots\dots (a_n, b_n)\}$ and unlabeled dataset with their pseudo labels $\{(\tilde{a}_1, \tilde{b}_1), \dots\dots (\tilde{a}_m, \tilde{b}_m)\}$ where $\mathbf{iter} = 1,2,3$:

   The following are the three variations of $\eta_{\mathbf{iter|curriculum}}$ used[1]
   1. MixUp
   2. Copy-Paste + Dropouts
   3. MixUp + Copy-Paste + Dropouts

   3.2. Use $\mathfrak{I}_s^{\mathbf{iter}}$ as a new teacher model to infer new pseudo labels $\{\tilde{b}_1, \tilde{b}_2 \dots\dots \tilde{b}_m\}$ as per step (2).

   3.3. Repeat step 3.1 to train the next iteration of student model

[1] With classical Noisy Student training, the student model is trained with noise $\eta_{\mathbf{iter|noisy}}$ = MixUp + Copy-Paste + Dropouts where $\mathbf{iter} = 1, \dots. n$

4.4 Results

4.4.1 Datasets

The proposed methods were validated on the BraTS 2018 dataset. BraTS 2018 dataset consists of 285 patients in the training set, with four available modalities, T1, T1c, T2, and FLAIR. For the Noisy Student Curriculum Learning, the BraTS 2020 validation and training data which are not part of the BraTS 2018 training set are used as the unlabelled dataset, amounting to 209 patients. This information is contained in the name_mapping.csv file provided by the organizer. The intensities of the four modalities are z-score normalized. We have also run some experiments on BraTS 2020 dataset to determine the generalization ability of our methods.

4.4.2 Implementation details

The architecture and layers proposed in this study are implemented in PyTorch. To demonstrate the performance of CASPIANET++, we had implemented these layers in the baseline model DMFNet [90]. The DMFNET (Dilated Multi-Fiber Network) is an enhancement of UNET, and was chosen because it is efficient and lightweight. A poly learning rate policy is implemented where the initial learning rate is 0.001 multiplied by a factor $[1 - (e/\mathrm{E})^{0.9}]$ where e is the current epoch and E, the total number of epochs. Synchronized batch normalization is used with a batch size of 4. Training is set to 300 epochs with an image crop size of 128x128x128. We maintain the following data augmentations throughout all experiments: (1) centroid cropping on WT regions, (2) random flipping, (3) random rotation, (4) random intensity and scaling, and (5) elastic deformations. For the optimizer, we utilize loss as the sum of region-based Dice Loss and BCE.

4.4.3 Ablation studies on components of the CASPIAN blocks

We perform experiments to demonstrate the improvement of each of the components towards the final CASPIANET++ block. We also include an implementation of the SE block (DMF+SE) to determine the contribution of channel-wise attention in multiclass semantic segmentation. We evaluate the performance of the model in terms of the Dice Scores and Hausdorff distance reported from the BraTS 2018 online evaluation platform.

From the results in Table III, the DMFNET+SE achieved dismal Dice scores, suggesting that a channel-only approach was insufficient to provide improvements to the pixelwise segmentation tasks. Compared to DMFNET+SE, the spatial approach in DMFNET+SAAM led to a 3.812% gain in the WT region. While a channel oriented approach did not do well on its own, adding the CAAM to the SAAM (DMF+SAAM+CAAM) further improved the gains in all three regions.

Table III: Ablation studies on the effects of individual components of the CASPIAN blocks on Dice Score and Hausdorff95 - all models are trained for a total of 250 epochs for the BraTS 2018 Validation dataset

	Dice Scores (%)			Hausdorff95		
Model	ET	WT	TC	ET	WT	TC
DMFNET+SE	78.709	87.044	83.112	3.233	24.240	8.009
DMFNET+SAAM	79.419	90.226	84.493	2.855	6.691	6.405
DMFNET+SAAM+CAAM	80.229	90.633	85.314	3.079	17.667	5.664
CASPIANET	80.338	90.826	85.616	2.895	4.717	5.570
CASPIANET+MultiPlanar	80.922	90.945	84.457	2.733	4.726	5.526
CASPIANET+MultiScale	80.749	90.986	86.703	3.159	6.394	5.589
CASPIANET++	81.083	91.197	87.600	2.659	5.054	4.912

In the CASPIANET experiment, we introduced excitation to boost the values of the asymmetry masks which are closer to 1. Compared with the DMFNET+SAAM+CAAM, while the Dice Scores only showed marginal improvements of 0.1-0.3%, the Hausdorff95 distance showed marked improvements in all regions particularly in the WT region. This indicates that while the pixel accuracy did not change much, the predictions of the segmented regions are closer to the ground truth. Clinically, this would indicate that the shape of the segmentations more closely resemble that of the ground truth compared to the DMFNET+SAAM+CAAM.

Next, all multidimensional schemes are compared with a baseline, CASPIANET – to see how each variant performs individually and in CASPIANET++. First, we look at the multiplanar variant CASPIANET+MultiPlanar, which shows a 0.6% increase in the ET region. The extra symmetrical information, along all 3 plane views, falters in the TC region (-1.2%). On the other hand, the addition of multiscale information CASPIANET+MultiScale boosted the TC region (+1.1%). With the multiplane and multiscale combined, the CASPIANET++ essentially evaluates a multiscale CASPIAN block on three planar views, providing the network with a larger and more holistic spatial information of the tumor region. The results show a remarkable 1.98% improvement in the TC region from the CASPIANET architecture. When the multiplane and multiscale variants are ensembled as CASPIANET++, the overall scores achieved much better results compared to its individual components.

To illustrate the differences between the CASPIAN variants we present the predictions of sample BraTS18_CBICA_AAL_1 in Figure 12(b) – (f). The prediction by DMFNET uses the model provided by the authors of DMFNET [90]. Compared to the ground truth in Figure 12(a), DMFNET and CASPIANET appear to have over-segmentations (see red arrow in Figure 12 (b) and Figure 12 (c)).

The white arrows in Figure 12 (d) and Figure 12 (e) show where the multiscale and multiplanar versions were able to localise the shape of the smaller tumor structure much better than the CASPIANET in Figure 12 (c). Comparing the multiscale with the multiplanar version, adding the multiscale helped with the small tumor structure, but was not able to detect the slight protrusion in the larger tumor structure (see red arrow in Figure 12 (d)). In the combined version, CASPIANET++ (Figure 12 (f)), minor refinements to the outline of the tumor can be seen compared to their individual schemes (see red arrows in Figure 12 (d) and Figure 12 (e)).

For this sample, despite poor contrast between healthy and tumorous tissue, and severely imbalanced tumor classes, CASPIANET++ shows a 10.5% improvement Dice score in the WT region from the baseline DMFNET. Clinically, the improvement presented by CASPIANET++ in demarcating the tumorous region is highly beneficial for treatment planning for radiotherapy or surgical guidance in Gamma Knife surgery, as any over-segmentation will lead to the removal of healthy tissue that could adversely affect brain function.

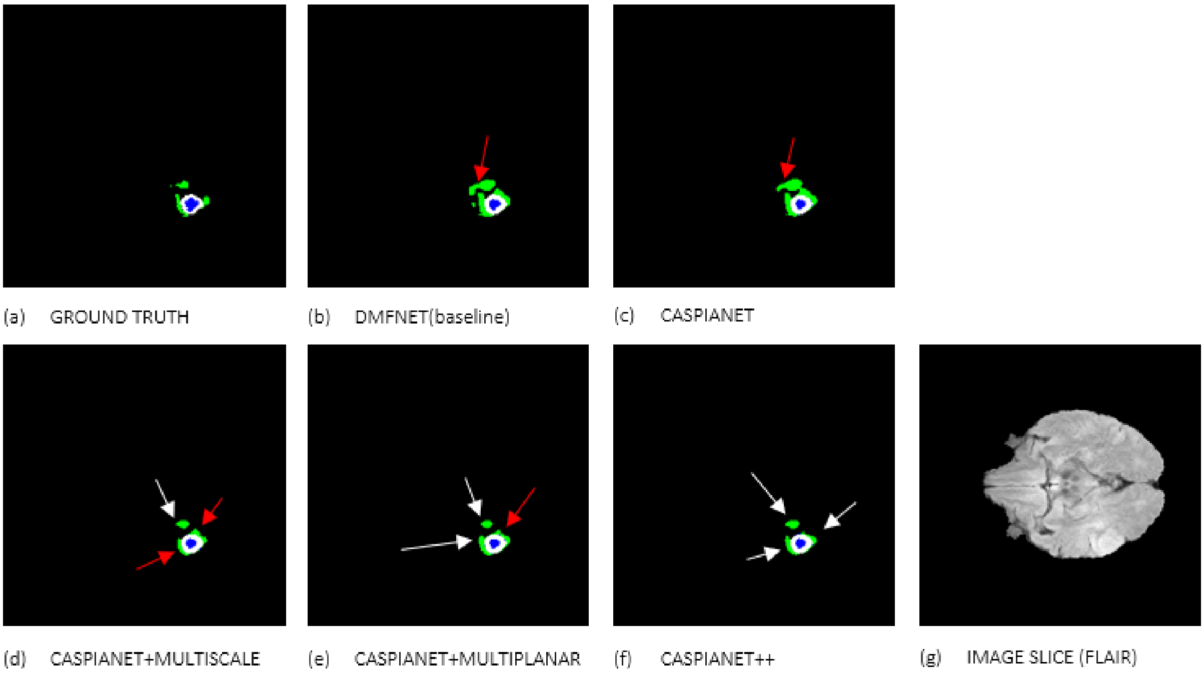


Figure 12 The series of images above depict the (a) ground truth, (b) predictions for baseline DMFNET model and (c-f) predictions for CASPIAN variants (g) FLAIR image for BraTS18_CBICA_AAL_1 (Slice 65). The Dice scores for ET, WT, TC are 85.003, 77.843, 91.413 (DMFNET) and 84.001, 88.354, 91.825 (CASPIANET++). White arrows denote regions of improved segmentation in each scheme, red arrows denote regions which did not perform well (Blue - Label 1, Green – Label 2, White – Label 4)

4.4.4 Visualization of heatmaps of the CASPIANET++ architecture

The heatmaps are generated from a selected feature map from the output of the last CASPIAN block of the CASPIANET++ architecture. As the feature map (120x120) is not spatially equal to that of the input image, a simple up-sampling is performed to produce the heatmaps shown in Figure 13. Regions of high heat intensity (depicted in red) correspond to regions of saliency. In general, heatmaps tend to follow the outline of the whole tumor region, with fairly accurate delineation of the boundaries. In the last row, the heatmap appears slightly smaller, due to the presence of small regions of high contrast artifacts in the opposite hemisphere, which the asymmetric mask will register as being similar.

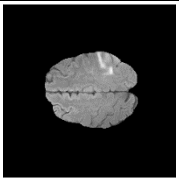
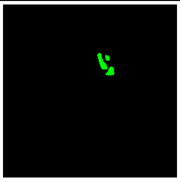
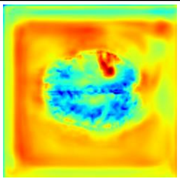
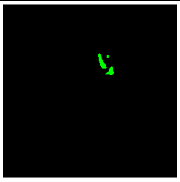
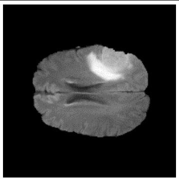
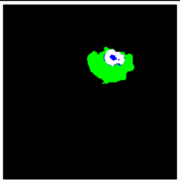
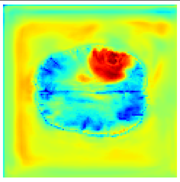
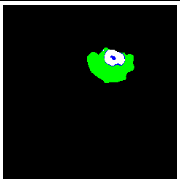
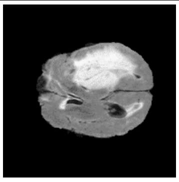
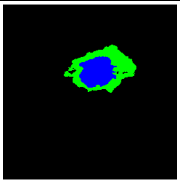
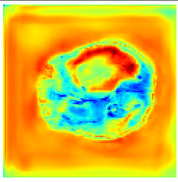
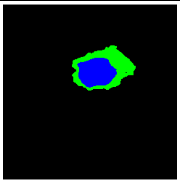
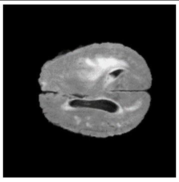
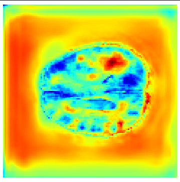
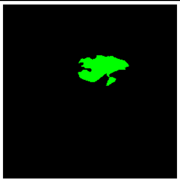
Patient data/Slice Index	(a) Slice data	(b) Ground truth	(c) Heatmaps	(d) Predictions
BraTS18_TCIA08_234_1/ Slice 110				
BraTS18_TCIA08_234_1 / Slice 86				
BraTS18_TCIA09_254_1/ Slice 78				
BraTS18_TCIA09_254_1/ Slice 88				

Figure 13: Image, Ground truth, heatmap and predictions from selected slices (indicated by slice index) obtained from patients BraTS18_TCIA08_234_1 and BraTS18_TCIA09_254_1 from the BraTS 2018 training set; The heatmaps in (c) are obtained from a selected feature map from the last CASPIAN block of the CASPIANET++, and the predictions in (d) are based on CASPIANET++ (Blue - Label 1, Green – Label 2, White – Label 4)

4.4.5 Comparison with other brain tumor segmentation methods

In Table IV we compare our validation results with the published results from the top 3 architectures from the BraTS 2018 competition. With the exception of CASPIANET++, all performance numbers were obtained from [90]. A PyTorch library available on github (https://github.com/sovrasov/flops-counter.pytorch) was used to obtain the performance values of the CASPIANET++ model. The library was validated with the reported numbers in the DMFNET paper before it was used to report on CASPIANET++. The winner of BraTS 2018, NVDLMED proposed an auxiliary VAE model to reconstruct the encoded image to regularize the main network [25]. The performance of this network is superior in ET regions. One possible factor is the spatial context provided by the large patch size (160x192x168) used during training compared to the other four architectures, DMFNET, No New-Net, CASPIANET++ (128x128x128) and DeepSCAN (5x192x192).

Table IV Results of BraTS2018 validation set obtained from the online validation tool and from published results.

	Dice Scores (%)			Hausdorff95			Performance	
Model	ET	WT	TC	ET	WT	TC	Number of floating-point operations ($x10^9$)	Number of parameters ($x10^6$)
NVDLMED [25]	81.73	90.68	86.02	3.8241	4.4117	6.8413	1495.53	40.06
No New-Net [94]	80.66	90.92	85.22	2.74	5.83	7.20	202.25	10.36

DeepSCAN [95]	79.60	90.30	84.70	3.55	4.17	4.93	-	-
DMFNet [90]	80.12	90.62	84.54	3.06	4.66	6.44	27.04	3.88
CASPIANET++	81.08	91.20	87.60	2.66	5.05	4.91	27.22	4.23

The authors of the baseline DMFNET trained their model with a batch size of 12 on 4 parallel Nvidia GeForce 1080Ti GPUs for 500 epochs [90]. We have achieved better scores using CASPIANET++ with a 50% reduction in the number of epochs and a 2/3 reduction in batch size on 2 parallel NVIDIA GeForce RTX 2080 Ti. The results in Table IV show that CASPIANET++ registers an increase in Dice Scores for all 3 regions over DMFNet. Our model appears to converge at a lower number of epochs, attributed by the CASPIAN/CASPIAN++ blocks being driven by CAAM and SAAM, which are a-priori cosine similarity calculations that do not require additional training.

The CASPIANET++ architecture introduces an additional 0.18G of floating-point operations (FLOP) and 0.35M parameters compared to DMFNET. NVDLMED, with its main and auxiliary branch, utilizes the most FLOP and parameters compared to other schemes. In contrast, CASPIANET++ uses 98% less FLOPs and 89% less parameters, while achieving modest improvements in Dice Scores for WT and TC regions and Hausdorff distance for ET and TC regions over NVDLMED.

To determine the statistical significance of these improvements, we use a Wilcoxon paired sample test on the results of the baseline model DMFNET and CASPIANET++. The test results show there is a statistical significance ($p<0.05$) in the improvement of Dice scores in all 3 regions (ET, $p < 0.0001$; WT, $p = 0.0010$; TC, $p = 0.0013$).

4.4.6 Noisy Student Curriculum Learning

In Table V, we compare the performance of Noisy Student Training [69] with our Noisy Student Curriculum Learning. CASPIANET++ showed marked improvements in the WT and TC regions, but the ET regions registered a 0.65% decrease in Dice Scores compared to NVDLMED. Noisy Student Curriculum Learning was proposed to boost the numbers in this region. The CASPIANET++ model reported in Table IV was identified as the teacher model to generate pseudo labels for the unlabeled dataset. Note that all experiments were conducted on the various training schemes using the same teacher model.

Table V Comparison of different self-training methods using CASPIANET++ as the Teacher and Student model. Unless indicated otherwise, each iteration is trained for a total of 300 epochs.

Model	Teacher-Student Training Iterations	Dice Scores (%)			Sensitivity (%)			Specificity (%)			Hausdorff95		
		ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
NVDLMED [25] (Without Self Training)	-	81.73	90.68	86.02	-	-	-	-	-	-	3.82	4.41	6.84
CASPIANET++ (Teacher Model)	-	81.08	91.20	87.60	82.42	92.51	86.59	99.81	99.43	99.83	2.66	5.05	4.91
Noisy Student Training	2	81.31	90.09	86.09	83.58	93.31	85.35	99.80	99.19	99.79	2.84	7.26	4.84
Noisy Student Curriculum Training	3	81.56	91.12	87.34	84.08	92.54	87.11	99.79	99.43	99.81	3.09	3.89	4.80

Table V shows that curriculum learning improves the ET region Dice scores of its teacher model from 81.08% to 81.56% (a 0.48% increase) after 3 iterations. Also, we managed to narrow the gap between the ET scores of NVDLMED to 0.17. In contrast, the Noisy Student Training at iteration 2 exhibits a 0.28% increase in the ET region, but experiences over a 1% decrease in the other two regions, compared to the teacher model. In addition, the noisy student training was erratic and could not predict any tumor regions in certain samples. This was not observed when curriculum learning was used. In comparing the Noisy Student with its curriculum learning counterpart, the specificity essentially remained the same, but the sensitivity increased for the ET and TC regions, indicating that the Noisy Student Curriculum method improved the quality of segmentation.

Noisy Student experiments were conducted by training one iteration of the student model for 300 epochs. The student model's validation result was checked every 10 epochs to select the best student model to become the teacher model in the next iteration. In the Noisy Student Training, we stopped at 2 training iterations as the scores for the WT and TC regions had degraded. This shows that in semantic segmentation problems, it may not be advantageous to introduce noise so early in the self-training iteration cycle. Experiments using stochastic depth to simulate noise in the model did not perform so well in CASPIANET++, possibly because the model is not deep enough to reap the benefits of dropping layers for regularization. Also, adding layers to increase teacher model capacity as suggested in [69] did not yield any benefits while increasing model complexity and training time.

4.4.7 Generalization ability of CASPIANET++ on the BraTS2020 validation dataset

Table VI presents the validation results obtained from the Noisy Student Curriculum Learning model to predict the BraTS2020 validation set. There was no further fine tuning to our models to obtain these results. The Noisy Student Curriculum model performed well, surpassing the TC region's Dice Scores from the Modality-Pairing method [96], which had placed second in BraTS2020. Compared with CASPIANET++ that was trained solely on BraTS2018 data, training with incrementally-noised data in the Noisy Student Curriculum Learning method enabled the model to generalize to unseen data. In contrast, the Noisy Student Training, which was trained with noise from the first iteration, did not perform as well.

Table VI Performance results for the BraTS2020 validation dataset from the online evaluation platform

	Dice Scores (%)			Sensitivity (%)			Specificity (%)			Hausdorff95		
	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Modality-Pairing	78.50	90.70	83.70	78.30	90.10	80.40	100.00	99.90	100.00	32.25	4.39	8.34
CASPIANET++	77.37	89.26	81.56	77.11	89.19	79.53	99.97	99.82	99.92	27.13	7.22	9.45
Noisy Student Training	77.29	87.84	83.55	76.86	91.90	82.53	99.97	99.86	99.96	29.80	14.37	8.16
Noisy Student Curriculum training	78.49	90.62	84.26	78.76	91.58	84.08	99.97	99.90	99.95	23.62	6.06	5.20

4.5 Conclusion

In this chapter, we proposed a novel multiscale and multiplanar architecture called CASPIANET++ for accurate segmentation of brain tumors (81.08%, 91.20%, 87.60% respectively for ET, WT and TC), resulting in faster convergence, smaller batch sizes and less GPU resources and marked improvement over the base architecture [90]. This method is different from conventional attention which relies on stochastic methods. Instead, we used the inherent symmetry of the brain to distinguish anomalous regions. With multiscale information from their higher-level counterparts, the attention mechanism is provided with higher scale resolution, which may be lost when the encoder converts spatial to contextual information. To account for tumor non-homogeneity, multiplane information is included as spatial asymmetry from a single axis does not provide a holistic picture of the tumor regions, particularly at boundary regions. The result is a lightweight attention block which combines a-priori information with excitation to amplify strong regions of dissimilarity and suppresses non-pertinent information.

ET regions have always registered the lowest Dice scores. In efforts to improve the Dice score in the ET region, we used Noisy Student training. This method was first used in classification of everyday objects – but we ran experiments to see how it would fare in a multiclass brain tumor segmentation problem in a low data regime. The Noisy Student improved the ET scores at the expense of lower scores in the WT and TC regions. This suggests that bombarding the network with noise early in the self-training iterations destabilizes the network, preventing it from generalizing. Also, the drop in the WT and TC scores in the first iteration may be symptomatic of catastrophic forgetting. In the original Noisy student method [69], which led to better ImageNET classification, strong emphasis was placed on using as much noise as possible from the first iteration of the student model. For brain tumor segmentation tasks, the host of noising techniques applied at the onset may have led the student model to “forget” the knowledge from the relatively noise-free teacher model. Hence, a new curriculum learning paradigm was introduced to Noisy Student learning by incrementally adding noise to the training samples, leading to better scores in the ET region (81.56%, 91.12%, 87.34% respectively for ET, WT and TC). This observation is reiterated by the BraTS2020 validation performance. Our results from the Noisy Student Curriculum learning paradigm suggest that incrementally introducing noise to iteratively train a segmentation network with scarce data may lead to gains in accuracy for other tumor segmentation problems.

Our results therefore suggest the CASPIANET++ architecture and the Noisy Student Curriculum Learning can be applied independently or combined to form a two-prong approach to segment tumor regions in medical images in the presence of small annotated and unannotated datasets. As we have demonstrated, the CASPIAN++ layer can be incorporated into any architecture to improve accuracy with minimal impact to computational requirements. The results can be further enhanced by applying a Noisy Student Curriculum Learning to unlabeled data. Here, the flexibility of the “curriculum” allows the method to be adapted to the problem. In our future work, we plan to determine how different definitions of “curriculum” can be used to train the student model to further improve accuracy.

5 CAVEAT: Gradual Self Training via Confidence and Volume based Domain Adaptation for Multi dataset Deep-Learning based Brain Metastases detection using Non-Local Networks on MRI images

5.1 Key points

- Lung nodule detectors can be used to detect brain metastases, as they are quite similar in appearance
- Non-local blocks improve the deep learning model through the addition of information not captured by a local convolutional operation
- The CAVEAT method gradually increases confidence scores and volume of brain lesions during training

5.2 Introduction

In the previous chapter, the focus was on the BRATS dataset which are primary gliomas. Morphological indicators and training with pseudo labels by progressively increasing noise in self-training were shown to be effective in the segmentation of primary gliomas. However, the SUNMET dataset consists of metastatic brain tumors. The tumors are smaller in size and do not have clear blood brain barrier compared to gliomas. While the existing system was trained primarily on gliomas, work needs to be done to adapt to the task of detecting metastatic tumors. It is imperative that the system detects small tumor structures as they are difficult to annotate and for targeted treatment via Gamma Knife Therapy.

Because majority of work on brain tumor segmentation are dominated by BRATS, only a handful of deep learning methods are developed for brain metastases detection. By using the CASPIANET++ model on SUNMET data, the system did not infer any brain lesions. We postulate that the large glioma structures cause more distortion in the symmetry of the brain than metastases.

The scarcity of fully annotated medical images in SUNMET makes the task of tumor detection a good candidate for self-training. However, self-training methods are primarily carried out on image classification of non-medical datasets using ImageNET. For the SUNMET detection task, self-training methods are developed by incorporating SUNMET with BRATS to result in a more universal tumor dataset.

In this section we aim to:

1. Detect BM using a 3D CNN based object detection method through efficient use of non-local networks through a single channel network using data from various MRI sequences
2. Develop a method to leverage specific characteristics in data for domain adaptation via gradual self-training to prevent catastrophic forgetting
3. Improve the performance (increase sensitivity and reduce FP) of the BM detection network by incorporating a multi dataset strategy, and leveraging out-of-domain data from public datasets

5.3 Convolutional Neural Network (CNN) Details

5.3.1 NLMET: Non-Local BM Detection Network

The model was based on N-Net, a lung cancer detector for Computed Tomography (CT) images that won the Data Science Bowl Lung Cancer Detection Competition [38]. Due to the parallels of size and multiplicity in lung nodules and BM, our work used the competition winning lung nodule detector architecture as the basis for BM detection [38]. The proposed model was implemented on Pytorch framework (v1.4.0, Meta AI, New York City, New York), Python (v3.8.5, Python Software Foundation, Delaware, United States), on a standard PC with 2 NVIDIA GeForce RTX 2080 Ti (Santa Clara, California).

The backbone of N-Net contained a UNET-like structure with 4 down-sampling blocks and 2 up-sampling blocks. The last 2 blocks (see

Figure 14(b)) were replaced by a Region Proposal Network (RPN), consisting of two CNN layers with 1x1x1 kernels, which produced an output of 24x24x24x5x3. This output tensor represents every location in 24 x 24 x 24 with 3 anchors of sizes 10,15 and 20 and the 5 regression values, corresponding to the probability, coordinates in x,y,z and the length of the bounding box. The probability is an objectness score provided by the RPN, which represents the degree of certainty which the location contains a metastases. This objectness score will be used in pseudo-labeling of the BrainMetShare dataset in our gradual domain adaptation method.

To increase depth in the UNET, residual units from [97] were introduced in the down-sampling path (see

Figure 14(c)). Each convolution layer was followed by batch normalization. The original N-Net architecture used for lung nodule detection consisted of an input block with coordinate locations of the input crop in relation to the position of the lung [38]. This location information is on the pretext that malignant lung nodules are typically found at the edge of the lungs. In NLMET, this we omit the location block because brain metastases can be found anywhere in the brain [98].

Due to GPU memory limitations, we reduced the crop size from 128 to 96 with a batch size of 4. However, smaller crop sizes tend to demonstrate lower accuracy in various CNN architectures [99]. To overcome this limitation, we integrated a non-local block at the skipped connection. In the UNET, these skipped connections compensated for the loss of spatial information during deconvolution. However, skipped connections utilize feature maps in early layers which do not have sufficient contextual information for differentiation of background and regions of interest within the volume. Non-local blocks were retrofitted into the skipped connections to evaluate attention in volumetric MRI data (see

Figure 14(a)). For computation efficacy, our method placed the non-local block only at the first skipped connection to provide contextual information to the up-sampling path (see

Figure 14(b)). A subsampling trick was also applied on the non-local blocks via max-pooling layers to reduce computational complexity. In

Figure 14(d), we have proposed another placement of the non-local block at the skipped connection to compare the performance of the two configurations. Our model is called **N**on-**L**ocal BM N**et**work, NLMET for short.

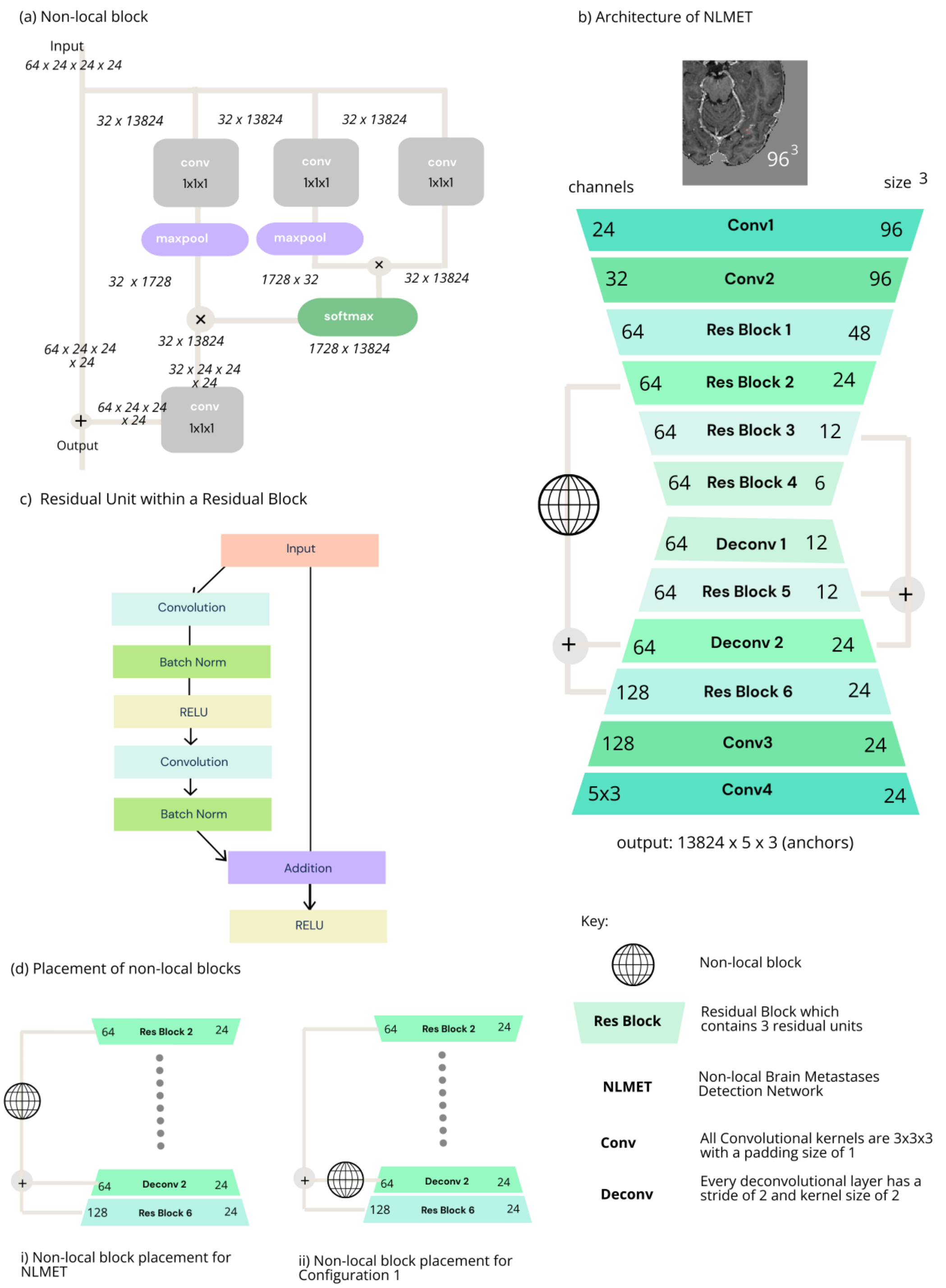


Figure 14 Architecture diagram of major components (a) Non-local Block, (b) NLMET, (c) Residual Units within a Residual Block, of the NLMET BM Detection Network, (d) Placement of Non-local blocks in NLMET and Configuration 1

5.3.2 Training details

The previously defined three datasets were used for training the network. Prior to training, all images were skull stripped by applying the brain mask generated by the BrainMaGe package [100]. All non-zero intensities below 0.2% and above 99.8% quantiles were clipped. All images in the datasets were then normalized to a [0-1] range via min-max normalization. To transform the BM segmentation datasets for use in BM detection, the coordinate of the centroid of the tumor is obtained through 3D connected component analysis of the ground truth mask (see Figure 15). A lesion was described as a cube with length equivalent to the longest lesion size.

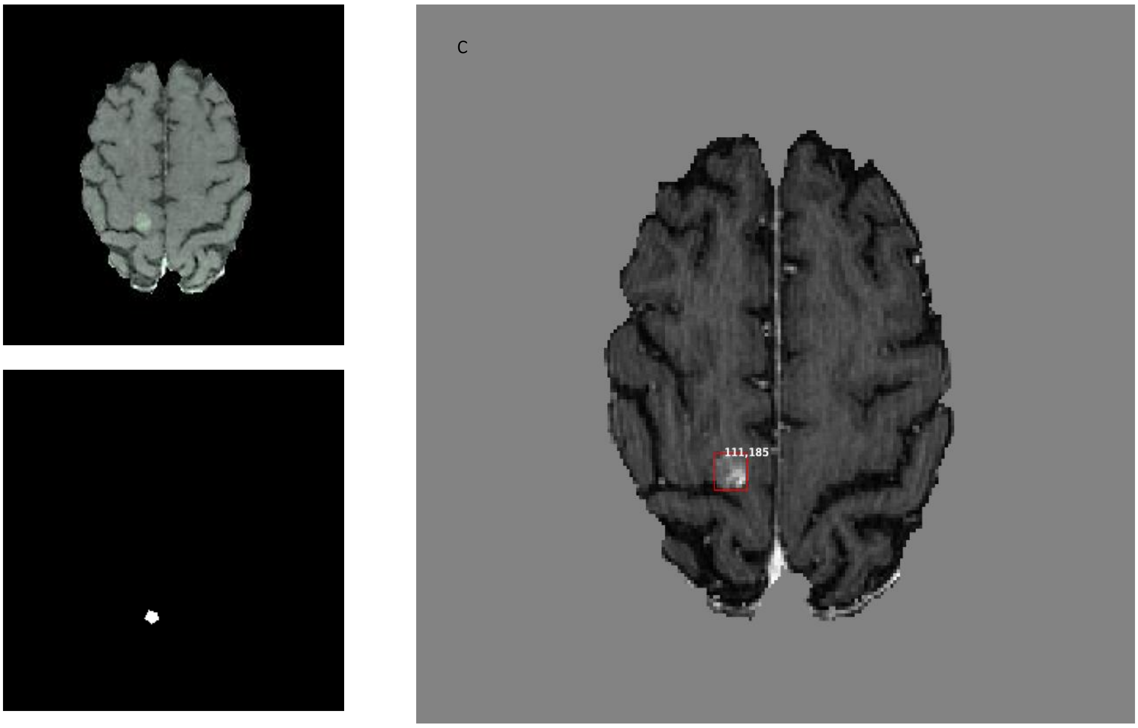


Figure 15 Representative axial slice of a brain MRI study (a) Original image from T1 spin-echo post-contrast sequence (b) corresponding ground truth provided by BrainMetShare and (c) resulting bounding box with coordinate locations of the tumor centroid generated by connected component analysis and considered as ground truth for the detector network on the post-processed image

During training, patches of size 96x96x96 were sampled randomly within the centroids of the lesions. We omitted sampling negative patches as the background area (negative samples) outnumber the lesions found in a training patch. Additionally, negative hard mining was implemented to only select the top k negative samples to be considered in the loss function. In our work, we used $\mathrm{k} = 2 * \mathrm{batch\_size}$.

An adaptive learning rate was used with an initial rate of 0.001 for the first half of the epochs, reduced to a factor 0.1 of the initial rate till the 80th percentile of epochs and finally to a factor 0.01 for the remaining epochs.

5.3.3 Loss function

Let the ground truth of a training input patch be defined as $\mathrm{T} = [\, \mathrm{t_x}, \mathrm{t_y}, \mathrm{t_z}, \mathrm{t_d} \,]$ where there are N instances of ground truth lesions. The output from NLMET describes a bounding box for an anchor, $\mathrm{B} = [\; \mathrm{b_p}, \mathrm{b_x}, \mathrm{b_y}, \mathrm{b_z}, \mathrm{b_d} \,]$ where $\mathrm{x}, \mathrm{y}, \mathrm{z}$ are coordinate locations and d is the bounding box length (or lesion diameter) and its corresponding probability, $\mathrm{b_p}$.

The first part of the loss function $\mathrm{L_{MSE}}$ is a mean-squared error loss, only the lesions are considered. $\mathrm{L_{MSE}}$ is the L2-norm between the bounding box coordinates and the ground truth, where N is the number of ground truth lesions, defined as:

$$\mathrm{L_{MSE}} = \frac{1}{\mathrm{N}} \sum_{1}^{\mathrm{N}} (\mathrm{t_x} - \mathrm{b_x})^2$$

The second loss, $\mathrm{L_{BCE}}$ is the binary cross entropy of $\mathrm{b_p}$ and its ground truth $\mathrm{g_p}$. Note that $\mathrm{g_p} \in \{0,1\}$ where 0 is background samples and 1 is for lesion candidates.

$$L_{BCE} = g_p \log(b_p) + (1 - g_p)\log(b_p)$$

The loss function L_{box} for each anchor box is the sum of both losses defined as:

$$L_{box} = L_{MSE} + L_{BCE}$$

5.3.4 Inference during testing

Due to GPU memory constraints, it was infeasible to obtain the output using a single pass of the network. Instead, the image was split into 96x96x96 patches in a sliding window fashion – with a window of 32 pixels. This translates to a maximum overlap size of 64x64x64. The output (x, y, z, r, p) from all the patches will result in proposals. The combined patches are processed with non-maximum suppression (NMS) to remove overlapping proposals.

5.3.5 Statistical Analysis

The model's capacity to detect metastases was determined by measuring the detection sensitivity and FP rate per patient (FP/patient). Free receiver operating characteristic (FROC) curves were generated at 7 predefined FP rates per patient: 1/8, 1/4, 1/2, 1, 2, 4, and 8. To ensure consistency, scripts were adapted from those provided to evaluate lung cancer detection in the LUNA Challenge [101]. This script was implemented on the Python package (v3.8.5, Python Software Foundation, Delaware, United States), with Scikit-learn (v1.0.1), Matplotlib (v3.5.1) libraries.

To determine the impact of the non-local block on the original model's performance we compare the sensitivity and FP/patient of NLMET with the original N-Net. To compare the sensitivities of both networks, we use a FP/patient of approximately 4.

In this work, we are presented with datasets with limited sample size (the ground truth from the BrainMetShare Validation dataset is not accessible), therefore we employ a 5-fold Cross Validation (CV) method to evaluate the overall performance of NLMET. The data is partitioned into five equal folds; 4 folds are used to train the network while the remaining fold is retained as the test set. Training is performed on four folds, until each fold has been used once as the test set (in our case, training is repeated 5 times). The average sensitivity and FP rates are taken from the 5 test folds. The train-test data split is generated using Scikit-learn (scikit-learn v 0.24.2) k-fold function to eliminate bias.

The CAVEAT method is tested on the test fold with the median sensitivity and FP rates within the 5 test folds, instead of re-running the experiments in a 5-fold method for every iteration due to time constraints.

Consider a candidate lesion l in a 3D volume described by coordinates (x_l, y_l) at slice z_l within a cube with length d_l. Consider an annotated metastasis from the ground truth g given by $[x_g, y_g, z_g, d_g]$ where (x_g, y_g) are the coordinates of the centroid location of bounding box with length d_g at slice z_g.A candidate lesion was a true positive with respect to ground truth g when the condition in (1) was satisfied:

$$[x_g - x_l]^2 + [y_g - y_l]^2 + [z_g - z_l]^2 < d_g^2 \;\therefore l \text{ is a true positive} \tag{1}$$

To test for significance of the CAVEAT method and NLMET architecture, every sample from the BrainMetShare and the local validation cohort (total of 44 samples) is compared using the Wilcoxon Signed-Rank test at p<0.05 (one tail). A right-tailed test is used to ensure that there is increase in the Sensitivity value using the methods we have implemented.

5.3.6 CAVEAT: Gradual Self Training through Confidence and Volume Based Domain Adaptation

Here, the methods used to co-train the datasets will be described. We present three datasets for training NLMET, consisting of two BM datasets and an auxiliary dataset containing primary brain cancer patients. The first BM dataset contains labeled data from the SUNMET. In the second BM dataset from BrainMetShare, we utilize labelled data from the training set and also unlabeled data from its validation set. Lastly, we utilize a primary brain cancer dataset containing labeled data from BraTS.

In every batch, a 1:2:2 ratio of BraTS to BrainMetShare and Local samples were used. Amongst the three datasets, BrainMetShare and the SUNMET contained BM that were closer in domain compared to BraTS. The latter dataset consisted of primary gliomas, which are larger in volume than BM. Here, we assumed that lesion volumes of primary gliomas correlated with the appearance of metastases. The lesion volumes are determined by the counting the labels 1 and 2 (edema marked with label 4 was excluded) from the segmentation mask ground truth. A total of 375 BraTS samples with lesions of comparable volumes to BrainMetShare and the SUNMET were selected as auxiliary data to generate more datapoints for training NLMET. To train on the multiple datasets, we employed a two-prong strategy, one aimed at generating higher confidence pseudo labels in the BrainMetShare validation set and the other to partition the BraTS data through lesion volume to create mini datasets. For higher confidence pseudo labels, we utilized the objectness score from the output of NLMET at successively higher probabilities in every iteration. For BraTS data, we gradually increased the volume of the tumors in the training set. We call this method CAVEAT (**C**onfidence **a**nd **V**olum**e** Based Domain **A**dap**t**ation) (see Figure 16).

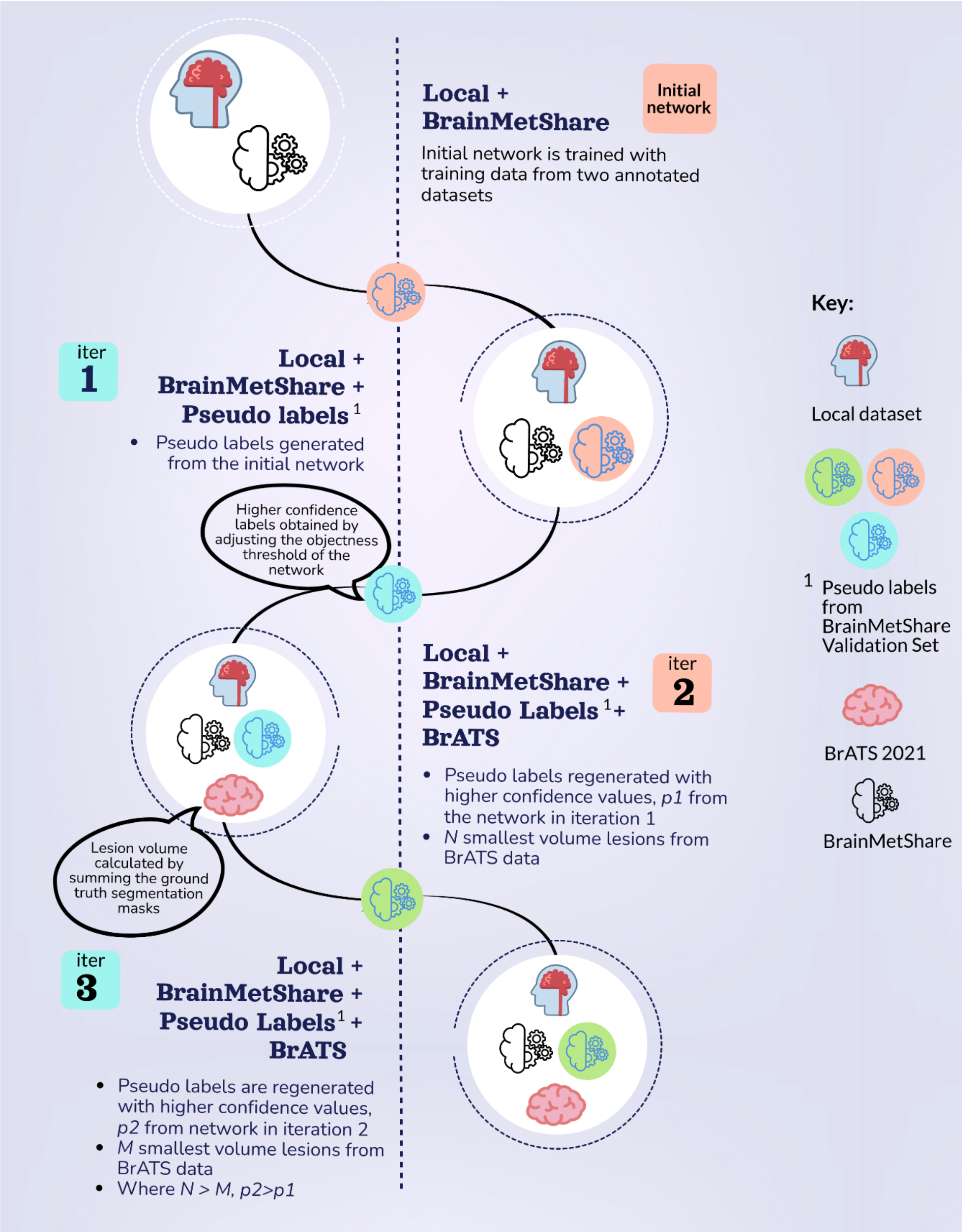


Figure 16 Flowchart description of CAVEAT. Pseudo labels were generated with successive levels of confidence and lesions were selected based on volume

5.4 Results

We presented a SUNMET dataset with 121 patients with brain metastases, where most patients had cancers of unknown primary (CUP) as they are symptomatic patients who have secondary metastases discovered on their first scan (Table VII). In both BrainMetShare and the SUNMET, we found that most of the patients had between 1 to 3 cancers.

Table VII Patient Demographics of BrainMetShare, SUNMET and BraTS dataset*

	BrainMetShare dataset	SUNMET	BraTS dataset
Gender	Male: 51 Female: 105	Male: 56 Female: 65	Male: N/A Female: N/A Total: 1251
Dataset type:	BM	BM	Glioblastoma multiforme
Primary cancer:	Lung:99 Breast:33 Skin/melanoma:7 Genitourinary:7 Gastrointestinal:5 Miscellaneous:5	Breast:3 Pituitary:3 Lung: 12 Skin:1 Spinal cord:2 Colon:1 Unknown:99	Glioblastoma multiforme:1251
Number of lesions:	≤3: 64 (41%) 4–10: 47 (30%) >10: 45 (29%)	≤3: 85 (71%) 4–10: 27 (21%) >10: 9 (8%)	≤3: 1251 (100%)

*Numbers indicate the number of patients

Total training time for 150 epochs with batch size of 5 was around 29 hours on 2 NVIDIA RTX 2080 Ti (Santa Clara, California) GPUs. The inference time for a full MRI volume was 850 ms. In Table VIII, by comparing the N-Net model with NLMET and Configuration 1, we found that the addition of the non-local block improved sensitivity as well as FP/patient for models trained independently on the BrainMetShare and SUNMET. We found that placement of the non-local block at the output of the downsampling path yields better results than placing it at the upsampling path. The sensitivity and FP evaluated on the 5 test folds of the 5-fold Cross Validation (CV) method were used to determine overall performance of NLMET in our dataset (Table IX). It was found that BrainMetShare on the 5th test fold (CV5) and 3rd test fold (CV3) on SUNMET was the median sensitivity and FP amongst the 5 test folds, and were selected as the folds to be used for training and testing the CAVEAT method.

Table VIII Summary of the detection metrics for validation data on BrainMetShare and SUNMET for the base model, Configuration 1 and NLMET

Model description	BrainMetShare				SUNMET			
	TP*	FP*	FP/patient	Sensitivity	TP*	FP*	FP/ patient	Sensitivity
N-Net (base)	145	86	4.095	0.609	47	83	3.609	0.610
Configuration 1	166	83	3.952	0.697	50	78	3.391	0.649
NLMET (Ours)	170	74	**3.523***	**0.718**	50	76	**3.304**	**0.649**

*TP-True Positives; FP-False Positives *Numbers in **bold** denote the better model performance, higher sensitivity at lower FP/patient

Table IX Detection performance 5-fold CV on BrainMetShare and SUNMET on the NLMET model

	BrainMetShare						SUNMET					
	CV1	CV2	CV3	CV4	CV5	Average	CV1	CV2	CV3	CV4	CV5	Average
Sensitivity	0.737	0.688	0.525	0.766	0.727	0.689	0.678	0.683	0.610	0.543	0.585	0.620
TP	182	258	202	196	173	202.200	61	127	47	44	110	77.800
Missed BM	65	117	183	60	65	98.000	29	59	30	37	78	46.600
FP	93	92	93	96	97	94.200	105	107	93	115	107	105.400
FP/patient	4.429	4.381	4.619	4.571	4.429	4.486	4.375	4.458	4.429	4.792	4.458	4.502

*BM: Brain Metastases, *TP-True Positives FP-False Positives CV: trained models from the 5-fold cross-validation on the training set

Table X tabulates the results for the CAVEAT method of gradual domain adaptation incorporating all three datasets. We obtained BraTS results from iteration 0 and 1 that was trained only on BrainMetShare and the SUNMET, where the first iteration registered 0.638 at 2.200 FP/patient. In the second iteration, we found a drop in sensitivity, but at a lower FP/patient. It can be observed that the method generalized well to BraTS, with a sensitivity of 0.723 and a FP/patient of 2.240 at iteration 3 without any fine tuning. BrainMetShare performance also benefited from the CAVEAT method – with marked improvements of ~ 17% in Sensitivity with improved FP/patient as training progresses from iteration 0 to 3 (0.634 to 0.811). The CAVEAT iterations showed steady improvement from the fully supervised baseline model to the next 3 iterations where more confident pseudo labels and small volume BraTS data were introduced (see Figure 17 and Table XII).

Table X Brain metastasis detection performance of NLMET (Non-local Metastases Detection) when trained with the CAVEAT method.

Iter	DATASETS				BrainMetShare Validation				Local Validation				BRATS validation			
	Local	BrainMetShare		BraTS*	TP	FP	FP/p	SENS	TP	FP	FP/p	SENS	TP	FP	FP/p	SENS
	TR	TR	PS													
0	✓	✓	×	×	151	97	4.619	0.634	54	115	5.000	0.630	30	55	2.200	0.638
1	✓	✓	✓	×	171	73	3.476	0.718	60	89	3.870	0.779	27	44	1.760	0.574
2	✓	✓	✓	350	183	63	3.000	0.769	57	75	3.261	0.740	32	51	2.040	0.681

3	✓	✓	✓	375	193	62	**2.952**	**0.811**	57	72	**3.131**	**0.740**	34	44	**1.760**	**0.723**
Baselines					170	74	3.523	0.718	50	76	3.304	0.649	26	69	2.760	0.553

Iter: Iteration of the CAVEAT method, *Indicates number of samples used for training, TP: True Positives, FP: False Positives, TR: Training data, PS: Pseudo Labeled data, FP/p: False Positives per patient, SENS: Sensitivity, Baselines provide values when trained on the training set of its corresponding dataset ✓: Dataset used in training ×: Dataset not used in training, Numbers in **bold** denotes the highest sensitivity and lowest FP/patient achieved by CAVEAT method

Table XI Wilcoxon paired test results for Sensitivities for selected experiments conducted at p<0.05.

Experiments	p-value	Significance at p<0.05
Comparing original N-Net to NLMET	0.021	Significant large difference
Comparing NLMET trained on BrainMetShare and Local dataset with Iteration 1 of CAVEAT method	0.005	Significant large difference
Comparing NLMET trained on BrainMetShare and Local dataset with Iteration 2 of CAVEAT meth145od	0.012	Significant large difference
Comparing NLMET trained on BrainMetShare and Local dataset with final iteration of CAVEAT method	0.001	Significant large difference

Wilcoxon paired tests in Table XI shows that there is significant difference in the Sensitivities reported between the original model and incorporating the non-local block in the architecture. The next 3 comparisons show that each iteration shows significant improvement over its NLMET counterpart, trained with only ground truth data from the BrainMetShare and local dataset.

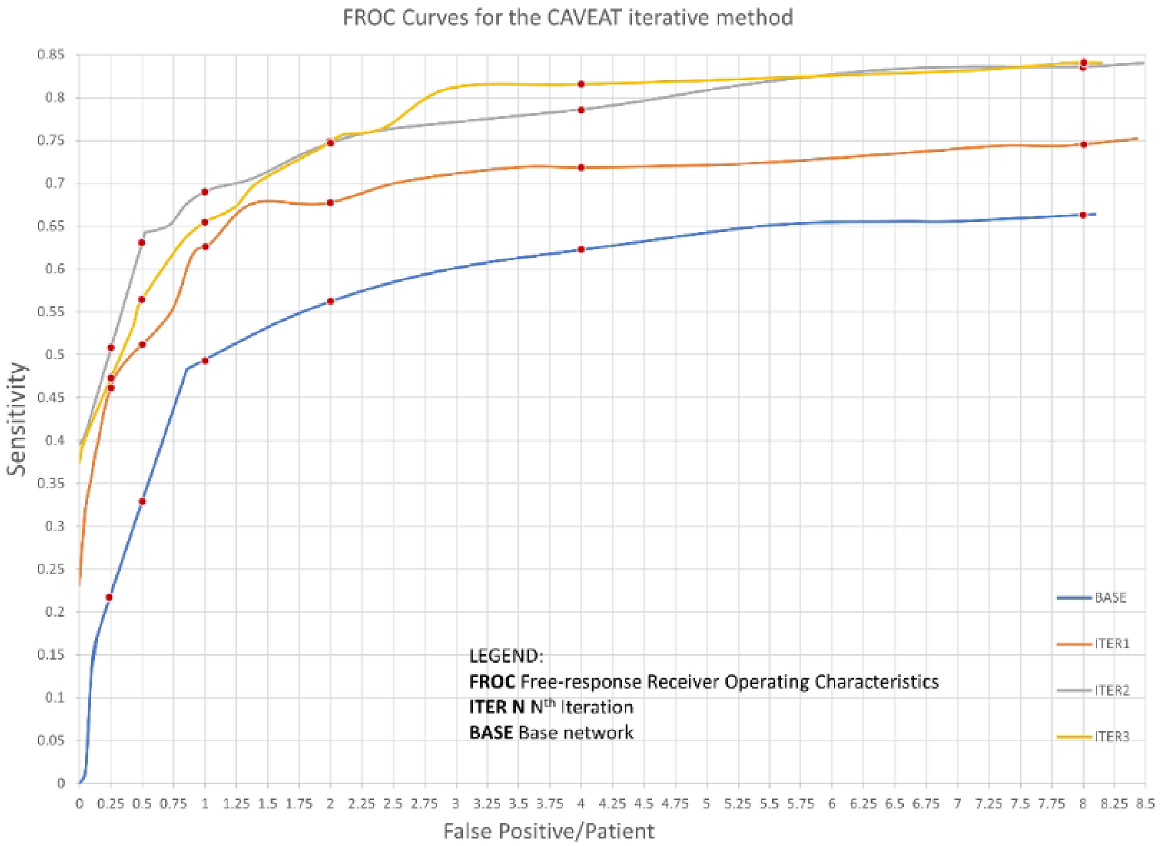


Figure 17 FROC curves for the CAVEAT method at different iterations evaluated at 1/4, 1/2, 1,2,4,8 FP/patient

Table XII Sensitivities at different iterations at 1/4,1/2,1,2,4,8 FP/patient based on the FROC curve in Figure 18

FP/patient	¼	½	1	2	4	8
BASE	0.218	0.332	0.499	0.569	0.631	0.671
ITER1	0.467	0.518	0.634	0.686	0.728	0.755
ITER2	0.478	0.571	0.663	0.757	0.796	0.846
ITER3	0.514	0.639	0.699	0.759	0.826	0.852

Correctly inferred lesions and false positives from the final iteration of CAVEAT can be found in Figure 18. These lesion candidates showed that CAVEAT helped the model detect different sizes of lesions. The FPs found in the representative subject of Figure 18c with motion artefacts were found to be blood vessels mistaken for metastases. A qualitative review of MRIs in the validation dataset found that similar FPs were also present in MRIs without motion artefacts (Figure 18d-e). Figure 18f is an example of the lateral ventricle mistaken as a FP. Missed detections were found to be common in samples with multiple tumors in close proximity, which were inadvertently removed by the NMS algorithm typically used in RPNs.

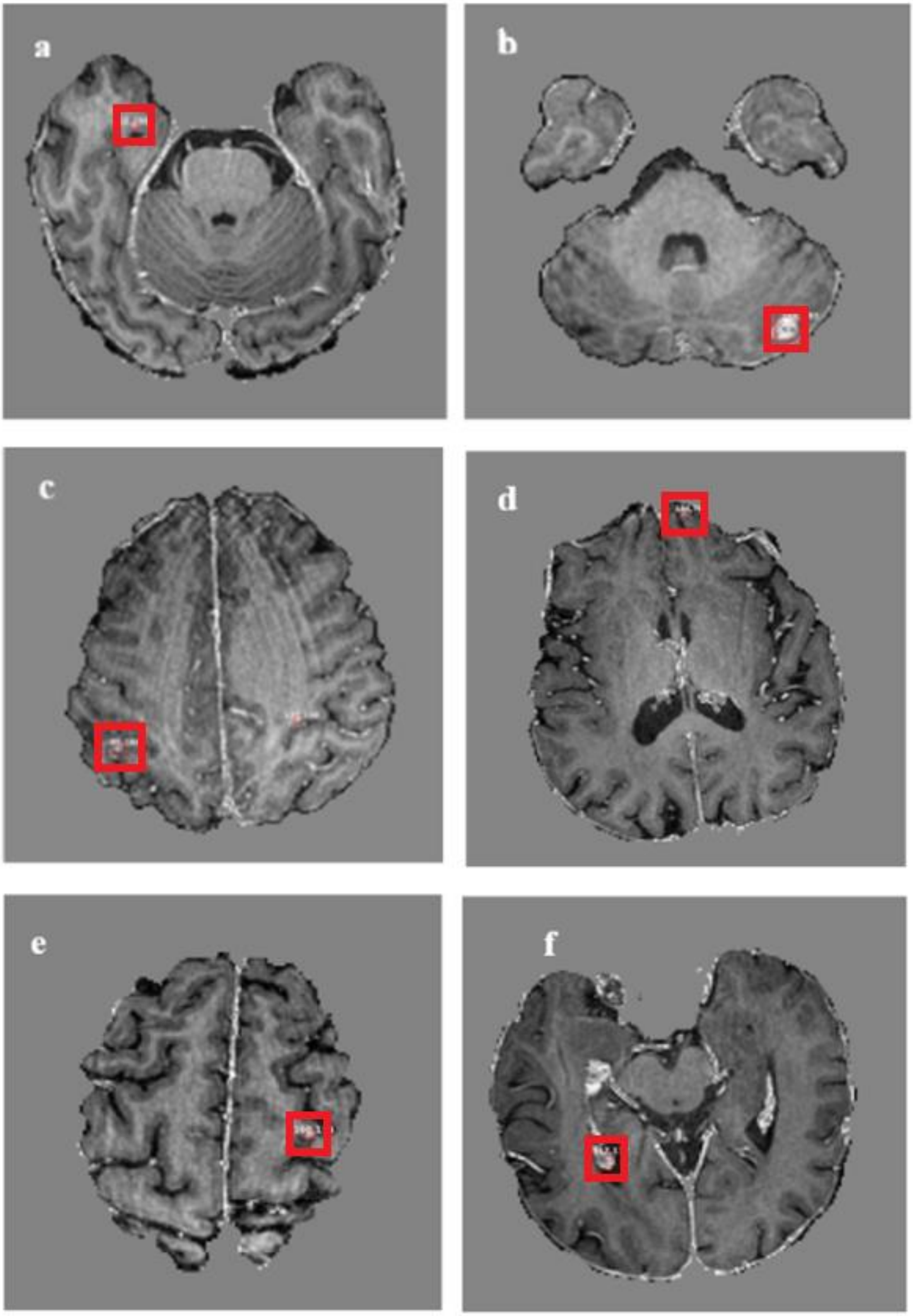


Figure 18 Detected lesions from the BrainMetShare dataset in six representative patients with lesions of varying sizes (a) small metastasis 2mm in size, (b) large metastasis, measuring 9mm (c) two false positives (d) one false positive (e) one false positive (f) one false positive

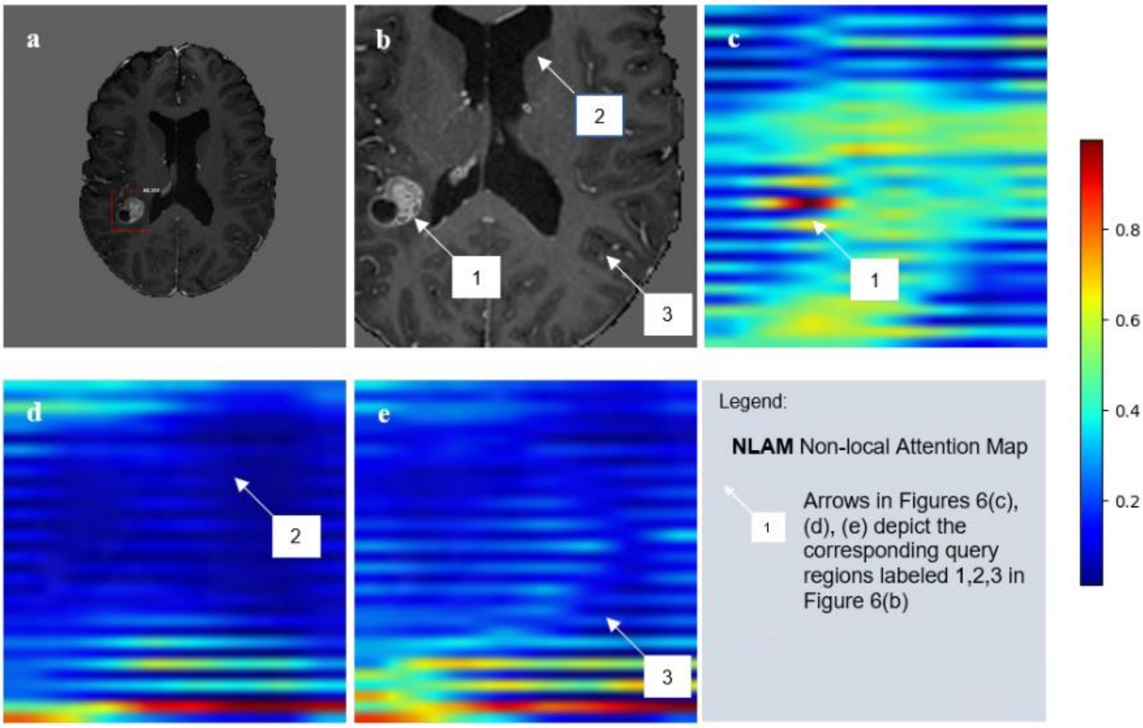


Figure 19 Depiction of the non-local response at 3 pre-determined locations using NLAM (a) Ground truth of tumor in a representative patient with 240x240 resolution (b) Patch of 128x128 which denotes the query regions of NLAM (c) NLAM near tumor region 1 (d) NLAM near non-tumor region 2 (e) NLAM near non-tumor region 3. The color bar correlates the colors shown on the heatmap to their NLAM values.

Figure 19 depicted heatmaps of three arbitrary points, with the point labeled 1 located at the tumor site, while points 2 and 3 are structures unrelated to the tumor. The purpose of the non-local attention maps (NLAM) was to compare the different non-local responses within different structures in the test image. While every point generated a volumetric NLAM, only a single slice was selected to be displayed. As observed from the NLAM in (Figure 19c-e), the response of each point in the test volume was different, and in the proximity of the tumor area (Figure 19c), regions of saliency are highlighted in red, whereas there is less response in areas that do not correspond with any tumors (Figure 19d-e). Artifacts in the form of horizontal lines observed in NLAM is attributed to the subsampling trick described in [52] to save computational resources. In our case, the subsampling reduced the patch from 32x32x32 to 32x32x8 and hence reshaping the image to its original size of 128x128x128 may have caused some distortions.

5.5 Discussion

In this study, we developed a deep learning method, NLMET, which detects BM reliably on different post contrast MRI sequences (T1-MPRAGE Post-Contrast, T1-Gradient Echo Post-Contrast and T1-FLAIR datasets). The various sequences demonstrate the potential of training with a mixed bag of MRI sequences, to simulate multi-center scenarios.

Also, the appearance of tumors on the MRI are dependent on the origin of the primary malignancies [98]. Since 82% of the data in the SUNMET has unknown primary tumors, we are not able to ascertain which tumor morphologies are under- or overrepresented in the dataset.

We utilized a lung nodule detection network as there are parallels between metastases in brain and lung nodules. For instance, nodules are comparable in size to metastases, and often mistaken for vessels and other structures, similar to how metastases are misrepresented as blood vessels in the brain. Additionally, using CAVEAT iterative gradual self-training framework for co-training of the three datasets further improved the quality of the detection network. Our findings were validated on held out data with samples of different sequences originating from three separate datasets.

We found that the addition of a single non-local block improved the sensitivity and FP/patient with BrainMetShare and SUNMETs by running ablation studies between the NLMET and N-Net on both datasets. While it would be intuitive to increase the number of non-local blocks in the object detection network, diminishing returns upon increasing the number of blocks have been found [52]. This result may be explained by the fact that BM cause subtle morphological distortions and midline shifts within the brain structure, which may not be adequately described by convolutional operations. These distortions serve as visual cues radiologists look out for in locating BM. As the goal of any CAD system is to model intrinsic knowledge, we included a non-local block to evaluate information which may not be within range of a convolutional operation. The NLAM suggests that the response for different regions in an image is different and methods that generate single attention maps for the entire volume [50], [102] may provide insufficient context to the network.

Prior work on BM was done on private datasets which does not allow comparison of the various neural networks [12]–[14], [34], [37]. We utilized results from the publicly available BrainMetShare dataset to benchmark the performance of our method with other methods that used the same dataset. With supervised learning on 3D NLMET, we achieved an average of sensitivity of 0.69 at 4.5 FP per patient, compared to the original BrainMetShare architecture using a modified 2.5D GoogLeNet with 0.53 sensitivity at 8.3 FP/patient [14]. The benefits of 3D NLMET was two-fold, it corroborated the effectiveness of non-local blocks over increased network depth (of the GoogLeNet) in capturing long range dependencies [52]. Secondly, it demonstrated the benefits of a 3D over 2.5D scheme – despite reduction in crop size and batch size. The results were obtained by 5-fold CV because the ground truth was unavailable for the validation set of BrainMetShare.

While we have shown that a non-local block can provide more global context, the networks were trained on their respective datasets in a fully supervised manner. This meant that each model had to be trained individually and there was no knowledge shared between datasets. Since all three datasets represented a domain shift from each other, training the model with all labeled data in one go may lead to model instability. It has been shown that systematic training such as curriculum learning proposed in (26) makes the model impervious to label noise. However, in our case there is no objective way to distinguish 'easy' cases from their more 'difficult' counterparts. Therefore, it is more advantageous to narrow the domain gap between datasets. An incremental domain adaptation approach was shown to yield higher accuracies at the target dataset compared to direct domain adaptation (27). One challenge in our multi dataset system is in determining the intermediate domains to which we would incrementally train the dataset.

To achieve a more generalized model and to leverage other datasets, we co-trained the datasets in a gradual self-training paradigm known as CAVEAT, which is inspired by incremental learning and curriculum learning [70], [103]. With CAVEAT, the aim was to create small shifts in domain so that the network can adapt to the target domain. Therefore, in the first iteration of CAVEAT we applied fully supervised learning on the local and BrainMetShare dataset as it had the smallest shift in domain. Next, we applied pseudo labels from the unlabeled data in

BrainMetShare. As the BraTS data was a primary glioma dataset, it had a larger domain shift, hence was only present from the 2nd iteration. In each batch, one BraTS sample was presented to two BrainMetShare and two local samples to prevent catastrophic forgetting as we were dealing with three different datasets. We found that a marginal increase in the amount of data in the next iteration was enough to illicit a better sensitivity at lower FP per scan. Also, a sudden increase in BraTS sample size may cause an imbalance in the number of samples being trained in an epoch, potentially causing overfitting.

In terms of stability of the CAVEAT approach, we evaluate this based on the performance of 3 datasets, where the improvements are consistent across the board. Without training on BraTS data (iteration 0 and 1) we can attain a relatively acceptable result on the BraTS dataset. We do find that after the 3rd iteration, there are depreciating returns as we find that we no longer have enough small volume tumors to train from and the confidence of the pseudo labels has reached its peak.

Empirical work on larger BraTS ratios (2:2:2) have been shown to be detrimental to the performance of metastases detection. As the main goal of our detection was to identify the smallest metastases and the trickiest to determine manually – we did not apply lesion size thresholds when evaluating the results. Validation of results on CAVEAT was performed with a single fold on the datasets to save on computation time.

After 3 iterations on CAVEAT, we found that the BrainMetShare and SUNMET improved at least 10% sensitivity and decreased in FP/patient down to 2.9. Another approach of synthesizing large numbers of lesions using GAN (Generative Adversarial Networks) achieved similar increase of 4 to 7 FP per slice [104]. Simply put, a radiologist must contend with 3 extra false lesions on each slice, whereby a single patient's MRI would have 450 incorrectly marked lesions (for an MRI with 150 slices). In the two-step methods we have reviewed, "CropNet" [36] gives 85% sensitivity at 5.85 FP/patient and the Faster RCNN with RUSBoost [12] registers 96% sensitivity at 20 FP/patient. In comparison with CAVEAT, the two-step methods achieved higher sensitivities, albeit at higher FPs. The reason for lower sensitivities could be attributed to the data distribution of the datasets and MRI sequences selected for use during training. In BrainMetShare, the authors contend that the diversity in data makes it a more challenging dataset to work with [14]. Similarly, the SUNMET has a range of different BMs ranging from 2mm to 40mm, with different distribution of primary cancers. One downside of higher FPs is that it creates additional overhead and is the main reason for poor adoption of CAD in clinical settings [9].

The presence of motion artefacts in a representative subject from BrainMetShare is expected as motion artefacts were not an exclusion criterion [14]. We have also ascertained that FPs were not only limited to MRIs with motion artefacts, so we rule out the possibility that there is limited training data with motion artifacts. Another instance of FPs are from high contrast structures in the MRI such as lateral ventricles. To reduce the prevalence of these FPs, these FPs can be incorporated in the training set for further negative sample training.

The CAVEAT approach of introducing increasingly confident labels and out of domain samples in a controlled manner provided regularization to prevent overconfidence in the network during self-training, evident by the 2.9 FP/patient result. However, no marked improvements were observed from the 2nd to 3rd iteration as the pseudo labels were sufficiently confident after 3 iterations, and the number of small lesions BraTS data was insufficient to perform further iterations. This is typical in iterative learning where improvements saturate after the 3rd iteration [69].

5.5.1 Limitations

Three imaging sequences were selected (T1-MPRAGE, T1-Gradient Echo and T1- FLAIR) for training with the aim of demonstrating generalizability to various sequences. Further studies on other combinations of multi parametric sequences to improve detection performance should be performed. Also, the SUNMET that we curated does not have enough data points for a training and validation set as performed in BrainMetShare. Secondly, there is a concern that the ground truth of the SUNMET was annotated by a single radiologist – we try to mitigate this by introducing two datasets, BrainMetShare, which is publicly available and annotated by two specialists and BraTS, which is annotated by four radiologists. The other limitation is that the two metastases' datasets, may not be similarly distributed as most of the primary cancers in the SUNMET are CUP – and therefore domain adaptation methods and self-training with out of domain BraTS dataset were perused.

5.6 Conclusion

In this chapter we have demonstrated the CAVEAT deep learning detection ecosystem to alleviate the scarcity of training data in a single dataset with a small FP rate and improved sensitivity. The proposed method has the added advantage of being MRI sequence-agnostic which is useful in clinical applications. The focus of our work is to detect smaller lesion structures from 1-3mm as these provide opportunity for early treatment for better prognosis of BM – by leveraging on an architecture developed for lung nodule detection purposes. The introduction of the NL block provided additional context to the model allowing for higher sensitivities and lower FPs compared to the architecture without the NL block – by evaluating the overall contribution of the feature maps with respect to a single pixel.

The CAVEAT method was trained on the premise of curriculum learning based on volume of the lesions and increasing the confidence level of pseudo labels. Overall, the method performed well, but we see a reduction in performance gains at the 3rd iteration, presumably because the model has reached its optimal solution. As we increase the 'objectness' score to reduce the noise in pseudo labels, we could inadvertently remove some true positives, particularly at later iterations. Additionally, as we increase the number of samples with large lesions, the out of domain training data may outnumber that of the primary training data – which could lead to catastrophic forgetting.

6 PHYSCISION: Improving Brain Metastases Detection in T1-MPRAGE scans through modelling Short-Range Dependencies and Dataset fusion strategies

6.1 Key points

- Brain metastases detection methods suffer from false positives and data scarcity
- Short-range non-local blocks in the deep learning architecture reduce the propensity of false positives
- Dataset fusion strategy uses datasets from different domains to alleviate data scarcity

6.2 Introduction

In the previous chapter, NLMET, a deep learning-based brain metastases detection method was described. To improve contextual information, non-local blocks were incorporated into the RPN. Three brain MRI datasets were used in the CAVEAT paradigm to improve detection ability.

In this chapter, short-range information is added to improve the performance of the non-local blocks. There are a few deficiencies in the CAVEAT model, namely uncertainties in pseudo labels and the poor efficiency from repeatedly training the model. The new approach harnesses fully annotated public datasets together with a single pass training method.

Data fusion based on similarities will mutually reinforce the training set and produce better outcomes [10]. However, there is no way to adequately compare the similarities between three datasets of different organs, MRI sequences and their respective anomalies. By leveraging the physical attributes of each dataset instead of quantifying the similarities between three unrelated datasets, this technique attempts to reduce the domain gap between them.

This chapter will demonstrate that:

1. Detection performance of brain metastases improves via architectural improvements by introducing short range dependencies for non-local blocks.

2. In medical applications where dataset size is limited, it is possible to incorporate publicly available datasets during training for better performance demonstrated using two datasets.

3. Domain adaptation in primary organ MRI datasets of different pathologies can be achieved through data fusion by synchronizing the data distribution.

4. Domain adaptation in different organ MRI datasets of different pathologies can be achieved through data fusion using quantifiable physical attributes.

6.3 Material and Methods

6.3.1.1 Patient demographics

156 patients with metastatic brain cancers who had undergone Gamma Knife Therapy in Sunway Medical Centre between May 2017 and August 2020 had participated in this study. Out of the 156 brain MRIs, 121 patients were included in the dataset with at least one metastatic lesion. 35 patients were excluded due to one of the following reasons 1) No metastatic lesions were found in the scans 2) Missing post-gadolinium T_1-MPRAGE sequences. A total of 987 lesions were found on Post-gadolinium T_1-MPRAGE sequences interpreted by a clinical interventional radiologist with 10 years of experience on a Carestream (Carestream Health) workstation. The Sunway Medical Center Brain Metastases dataset will be collectively known as SUNMET (Sunway Medical Center Brain Metastases dataset)

6.3.1.2 MRI protocol

Examinations were conducted on a 3T scanner (Siemens MAGNETOM Skyra) in the HFS (head-first supine) position. The acquisition parameters for post-gadolinium T_1-MPRAGE sequences were as follows: matrix = 256x256x192; slice thickness = 0.90mm; pixel spacing = 0.90-0.93mm; TE = 2.32-2.49ms; TR = 1900-2000ms; TI = 900ms; flip angle = 8-9^{o}.

6.3.1.3 Additional public datasets for dataset fusion

To investigate the effect of joint training on out-of-domain datasets on brain metastases detection, we introduce two publicly available MRI datasets, BraTS, a brain glioma segmentation dataset [105] and Duke Breast Cancer MRI dataset [106]. For the BraTS dataset, 3D connected component labelling on ground truth segmentation masks with labels 1 (necrotic core) and 4 (enhancing tumor) denoting the tumor core was performed on the ground truth to obtain the (x,y,z,diameter) centroid of the lesions.

For the Duke Breast Cancer dataset, the ground truth was given by (row_start, row_end, column_start, column_end, slice_start, slice_end) and additional preprocessing was performed to convert it to centroid information. In terms of the MRI sequences, since the scans came from various centers, we randomly selected the MRI modalities to introduce some variance in the training data. Also, it is important to note that some of the image sequences were ordered differently based on the ImagePositionPatient DICOM tag. The images were normalized so that the start_slice is always at the lowest height, and this was verified manually by visual inspection of all the ground truths (see Figure 20). 12 samples were excluded from the onset as prescribed by [107], because of missing pre/post contrast images.

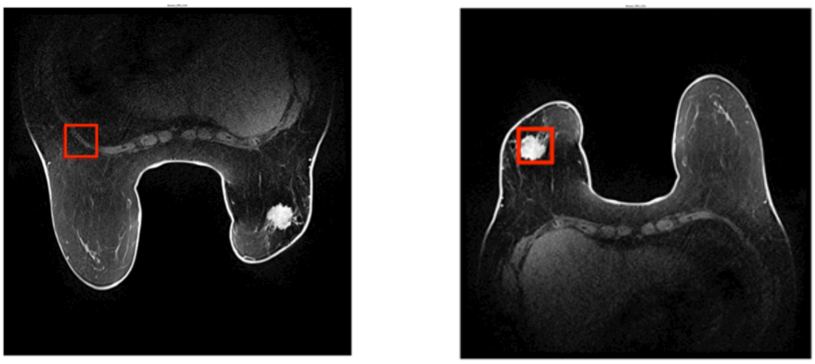


Figure 20: (a) shows the ground truth with start_slice at the highest z slice (b) after realignment, notice the ground truth is now situated within the lesion

6.3.1.4 Deep learning-based detection by modelling short-range dependencies.

The base of this model is a lung-nodule detector [38] which has achieved superior results in the Kaggle DSB competition. This detector features a Region Proposal Network (RPN) with ResNet blocks (see Figure 21(a)). Since, lung nodules and brain metastases share similarities in terms of appearance, we use this as our base model. However, authors in [38] contend that the model is not primed for smaller structures as lesions below 7mm are excluded from the training set. In our work, we do not exclude any lesions from the training set and introduce architectural changes to improve the detection ability of these smaller structures.

To enhance the detection ability of lesions, a modified non-local block (NLB) is introduced to improve the spatial information in the downsampling block. The original NLB calculates the dependency of each point relative to every point in the feature map [108]. Because distant frames are highly correlated with respect to the reference frame; the evaluation of long-range dependencies is crucial. When NLB is used in conjunction with convolutional blocks, it provides the long-range attention to complement the kernel operations, which act on smaller scale. This works well on video and temporal data, as video images are captured in multiple frames per second.

For brain lesions in MRI data, an average-sized lesion appears across 3-5 slices with minimal correlation between distant slices beyond 20 slices. Neuroradiologists typically rule out false positive blood vessels, by charting the paths of the structures for more than 20 slices. This suggests that NLB range is overextended, returning areas of saliency on blood vessels which contributes to false positives. By leveraging "soft knowledge" of experienced professionals, we introduce Short-Range dependencies for non-local (STROLL) block, which sufficiently describes metastatic lesions. Our experiments investigate 3 different placements of the STROLL block (see Figure 21(a)).

Training is done for 150 epochs, using an adaptive learning rate as follows:

$$\text{Current epoch, e out of total number of epochs E,}$$

$$\text{Learning rate denoted by } R_e \text{ is defined as}$$

$$R_e = 0.001 \text{ where } e < 0.5E$$

$$R_e = 0.1 * R_{e-1} \text{ where } 0.5E < e < 0.8E$$

$$R_e = 0.01 * R_{e-1} \text{ where } e > \ 0.8E$$

The model is trained on 96x96x96 patches with a batch-size of 4 on two NVIDIA RTX 2080 Ti (NVIDIA Corporation) GPUs. Given a 96x96x96 input patch will generate instances of ground truth:

$$G_{i,j} = [\, g_p, g_x, g_y, g_z, g_d \,]\Big|_{i\,=\,1\,\ldots.13842} \quad (1)$$

corresponding to 13842 x 3 anchor sets $j = j_1, j_2, j_3$ which describe centroid coordinates g_x, g_y, g_z and lesion diameter g_d with the g_p values:

$$g_{p=\text{lesion}} = 1 \qquad g_{p=\text{background}} = 0$$

Consider an anchor box $A_{a,j} = [\ a_p, a_x, a_y, a_z, a_d\ ]$ from anchor set j where a_x, a_y, a_z, are centroid coordinates and a_d is the lesion diameter and its corresponding probability, a_p

Consider L_{MSE}, the L2-norm between the centroid coordinates and the ground truth,

$$L_{MSE} = \frac{1}{I}\sum_1^I (a_c - {}_c)^2 \Big|_{c\,=\,x,y,z,d} \text{ where } I = 1 \ldots 13824 \quad (2)$$

While L_{BCE} is the binary cross entropy of the probability of the a_p and its ground truth g_p

$$L_{BCE} = g_p \log(a_p) + (1 - g_p)\log(a_p)\Big|_{p\,=\,\text{lesion, background}} \quad (3)$$

The loss function L_{box} for each anchor box is the sum of both losses defined in (2) and (3):

$$L_{box} = L_{MSE} + L_{BCE} \quad (4)$$

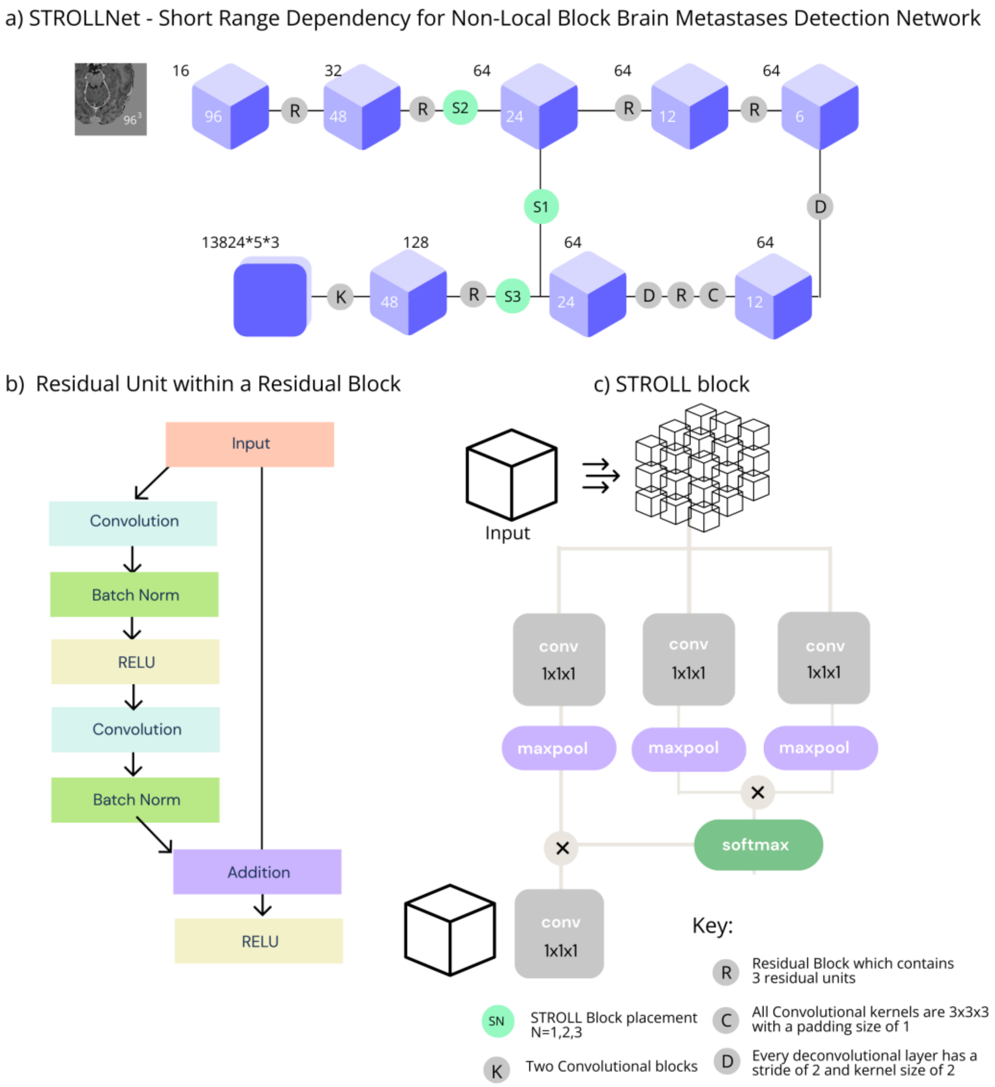


Figure 21: STROLLNet model architecture component diagram

6.3.1.5 Dataset fusion strategies

From the best model evaluated in the previous section, we peruse that model to further boost the data volume and their performance via training with dataset fusion strategies. We incorporate two out-of-domain datasets as auxiliary data to boost the sensitivity and reduce the false positive counts of the brain metastases detection. Because of the domain gaps in auxiliary datasets, we utilize quantifiable physical attributes and clinical factors to filter appropriate data points to achieve better performance in brain metastases detection. This method is called PHYSCISION (Physical attributes and clinical factors data fusion)

6.3.1.5.1 Dataset fusion of Same Organ, Different Pathology

BraTS primary glioma dataset presents an opportunity to investigate dataset fusion of another pathology within the brain. Gliomas and metastases appear similar in conventional MRI, which is an advantage for training the primary task [109]. Specifically for our dataset, Figure 22 depicts the tumor diameters for SUNMET and BraTS. Two observations can be drawn, 1) BraTS has considerably more tumors than SUNMET 2) SUNMET tumors are less than 64mm. To close the domain gap, we aim to create a mini dataset to mimic the distribution of the SUNMET dataset. However, curating a dataset according to tumor diameter is not straightforward when there are large and small tumors in proximity in a single patient. Therefore, the aim is to create a mini dataset which most closely resembles the data distribution of SUNMET. Two datasets, labeled D1 and D2 will be selected for use in experiments, by matching the distributions of tumor diameters to the SUNMET dataset (see Figure 24).

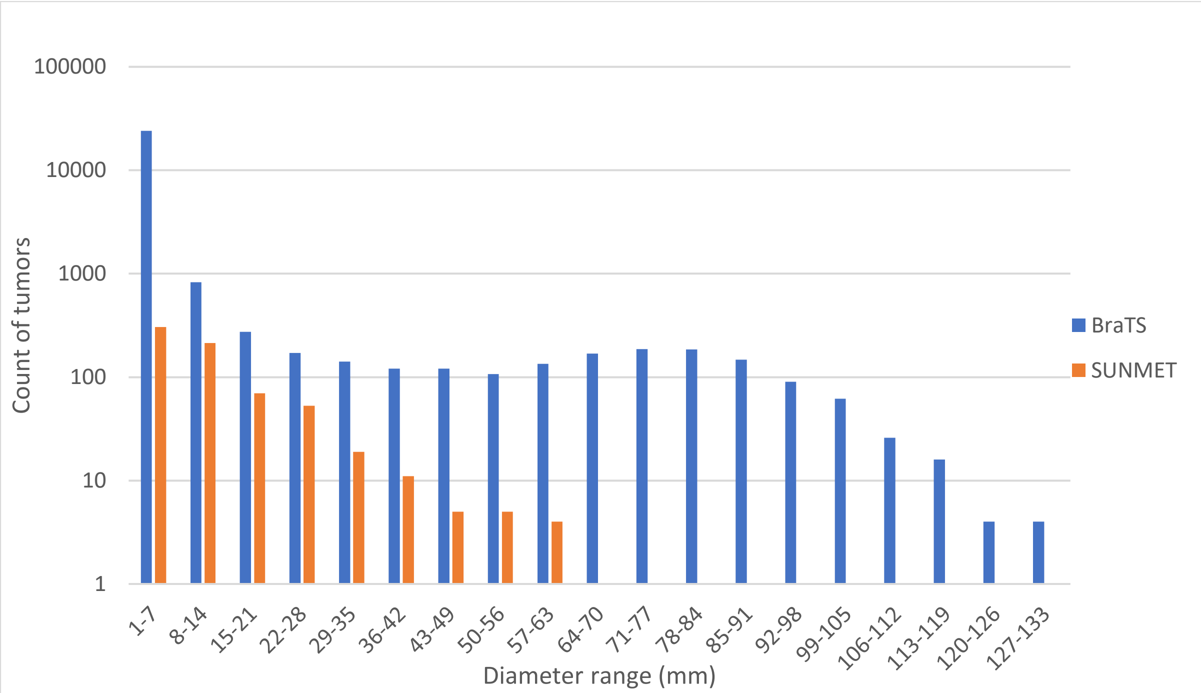


Figure 22: Distribution of tumour diameters in patient data from the BraTS and SUNMET dataset (note that y-axis is in logarithmic scale)

6.3.1.5.2 Dataset fusion of Different organ, Different pathology

The DUKE Breast Cancer Dataset showcases dataset fusion using a different organ and pathology. Since the dataset contains clinical factors as supplementary information, we examined Molecular type (MT) and Breast Density (BD) to create a mini dataset, instead of using tumor diameters as we did for BraTS data. As prescribed by curriculum learning [70], the groups are trained from one characteristic to another in an epoch. For BD characteristics, the model is trained from fattiest to the least fatty. For MT, we train them on a group basis from 0 to 3 (0 = luminal-like, 1 = ER/PR pos, HER2 pos, 2 = her2, 3 = trip neg).

6.3.1.6 Statistical Analysis

For consistency, results were generated from scripts provided to evaluate lung cancer detection performance in the LUNA Challenge [101]. This script was implemented on the Python package (v3.8.5, Python Software Foundation), with Scikit-learn (v1.0.1), Matplotlib (v3.5.1) libraries. The validation set was generated randomly via a 5-fold cross validation library to ensure randomness.

6.4 Results

6.4.1 Patient demographics

From a total of 170 patients, a total of 49 patients were excluded based on criteria in Figure 23. The final training and validation set consisted of 121 scans from 121 patients (age not available as some personally-identifiable information was redacted) who underwent Gamma Knife Surgery in Sunway Medical Center, Malaysia. Of this, 98 were used for training (643 lesions; size ± standard deviation 11.39 ± 10.31mm; range 2-83mm) and 23 for validation (76 lesions; size ± standard deviation 14.02 ± 10.51mm; range 3-60mm)

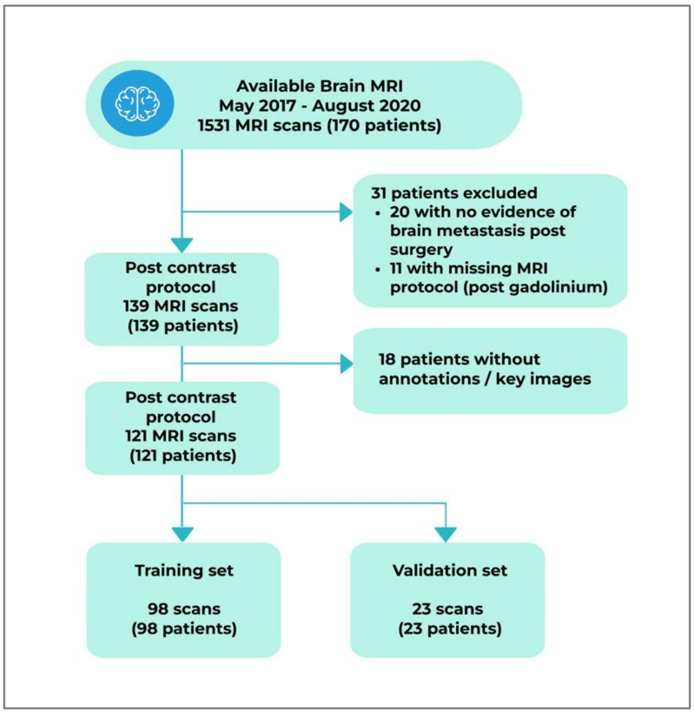


Figure 23: Flowchart of the data collection and selection procedure of the SUNMET dataset

6.4.2 Additional public datasets for training

To expand the SUNMET training set, we utilize two datasets from DUKE Breast Cancer Dataset and BraTS Glioma dataset outlined in Table XIII. During training, we sample 300 data points from the 1251 BraTS patients and around 150 samples from DUKE dataset to prevent the model from overfitting to the other two datasets.

Table XIII: Distribution of patients demographics in the BraTS, DUKE and SUNMET dataset (age information has been redacted for patient confidentiality)

	BraTS	SUNMET			DUKE
Variable	Overall sample	Overall sample	Training set	Test set	Overall sample
Total number of patients	1251	121	98	23	910
Gender distribution	Male: N/A Female: N/A	Male: 56 Female: 65	Male:42 Female:56	Male:14 Female:9	Female: 910
Primary cancer	Glioblastoma multiforme:1251	Breast:3 Pituitary:3 Lung: 12 Skin:1 Spinal cord:2 Colon:1 Cancers of Unknown Primary (CUP):99	Breast:3 Pituitary:2 Lung: 11 Skin:1 Spinal cord:2 Colon:1 CUP:78	Lung:1 Pituitary:1 CUP:21	Breast:910
Lesion sizes	≤7: 541 (58%) 8-12: 44 (5%) >12: 349 (37%)	≤7: 338 (3%) 8-12: 150 (51%) >12: 231 (46%)	≤7: 313 (49%) 8-12: 144 (22%) >12: 186 (29%)	≤7: 25 (33%) 8-12:6 (8%) >12: 45 (59%)	≤7: 0 (0%) 8-12: 24 (3%) >12: 882 (97%)

6.4.3 Deep learning-based detection by modelling short-range dependencies

In Table XIV, STROLLNET configurations S1 and S2 register improvements in sensitivity and FP. Compared with the original NLB, STROLLNET S2 logs an increase in sensitivity of 4.1%. Configuration S3, placed at the upsampling path did little to improve the sensitivity and FPs. All STROLL blocks were found to give better FP compared to their NLB counterpart. The best STROLLNET model, S2 was selected as the basis for all experiments on PHYSCISION.

Table XIV: Summary of the detection metrics for SUNMET validation data on various configurations of STROLLNET trained on SUNMET training data

Configuration	Performance metrics			
	TP*	FP*	FP/patient	Sensitivity
Non-local block (N1)	43	70	3.043	0.558
STROLLNET1 (S1)	45	68	2.957	0.584
Non-local block (N2)	42	65	2.826	0.545
STROLLNET2 (S2)	45	61	**2.652**	**0.584**
Non-local block (N3)	44	71	3.087	0.571
STROLLNET3 (S3)	44	63	2.739	0.571

*TP-True Positives; FP-False Positives *Numbers in bold denote the better model performance, higher sensitivity at lower FP/patient

6.4.4 PHYSCISION Dataset fusion strategies

Table XV is a comparison of the training performance of data fusion strategies. Experiment 1 demonstrates training on only SUNMET data. Comparing Experiments 2 and 3, BraTS dataset D1 yielded the better performance. Training with D1 increased sensitivity by 3.9% compared to training on the primary dataset. D1 yielded better performance than D2 and was used for all experiments involving DUKE datasets. Figure 24 shows the distribution of dataset D1 with fewer large lesions than D2, and through experimentation was shown to provide better results.

Table XV: Summary of detection metrics for SUNMET validation data on STROLLNET2 with data fusion strategies

Experiment	SUNMET	BraTS		DUKE			Performance metrics				Training time (hours)
	N patients*	N patients	Dataset Name	N patients	Attribute[1]	Categories	TP*	FP*	FP/patient	Sensitivity	
1	98	N/A	N/A	N/A	N/A	N/A	45	61	2.652	0.584	6.50
2	98	300	D1	N/A	N/A	N/A	48	58	2.566	0.623	16.50
3	98	300	D2	N/A	N/A	N/A	44	63	2.739	0.571	16.50
4	98	300	D1	910	N/A	N/A	46	56	2.435	0.597	67.50
5	98	300	D1	910	MT	0,1,2,3[2]	48	49	2.130	0.623	67.50
6	98	300	D1	910	BD	5,4,3,1,0,NC[3]	51	51	2.217	0.662	67.50
7	98	300	D1	37	BD	5,4,3[3]	53	51	2.217	0.688	17.00
8	98	300	D1	156	BD	5,4,3,1,0[3]	45	52	2.261	0.584	19.50
9	98	300	D1	83	BD	5,4,3,1[3]	**55**	**49**	**2.130**	**0.714**	18.75

[*] N patients – Number of patients
[1] MT – Mol Type, BD - Breast Density, N/A – Not applicable
[2] Mol Type (MT) Categories: 0 = luminal-like, 1 = ER/PR pos, HER2 pos, 2 = her2, 3 = trip neg
[3] **Breast Density (BD) Categories:** 0=heterogeneous, 1=scattered, 2=minimal, 3=moderate, 4=extremely, 5=predominantly fatty, NC=Not categorized

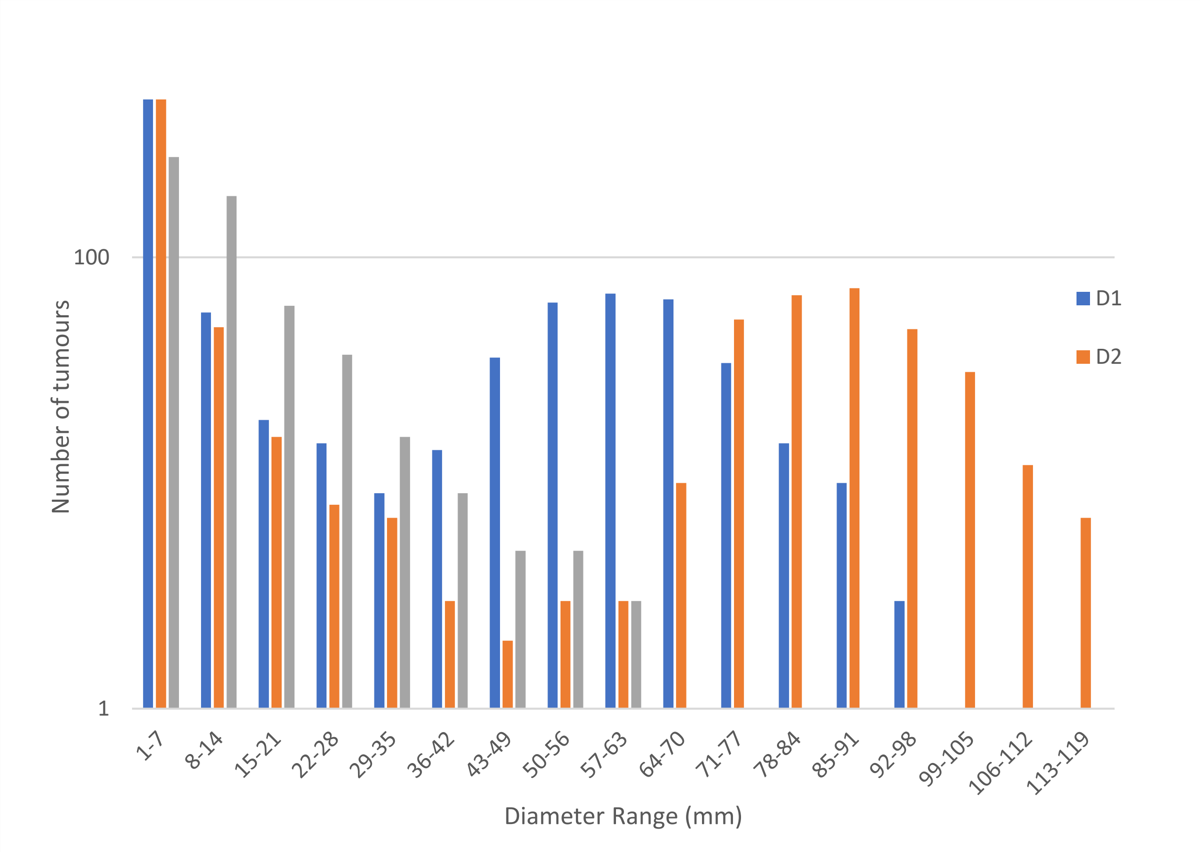


Figure 24: Tumor diameter distribution between two dataset BraTS mini datasets D1 and D2 and SUNMET. Datapoints in D1 and D2 are selected based on the lesion diameters and plotted in the graph to get a distribution as close to the primary SUNMET dataset. (Note that y-axis is scaled logarithmically)

Without any data fusion strategy, the best results came from combining the three datasets (see Experiment 4). Training of all 910 datapoints on MT versus BD characteristics, BD fared marginally better. The last 3 rows describe experiments performed on data with Breast Density categorized. When training on PHYSCISION, the dataset trained with 3 datasets on PHYSCISION, with only 83 datapoints from DUKE produced the best results (71.4% Sensitivity, 2.13 FP/patient).

Table XVI Wilcoxon paired test results for Sensitivities for PHYSCISION experiments from Table XV conducted at p<0.05.

Experiments	p-value	Significance at p<0.05
Comparing Experiment 1 and Experiment 9	0.004	Significant large difference
Comparing Experiment 1 and Experiment 2	0.198	Non-significant medium difference
Comparing Experiment 6 and Experiment 9	0.013	Significant large difference

From Table XVI, Wilcoxon paired tests were conducted to determine the significance of selected experiments. Interestingly, experiments with only SUNMED data and dataset D1 did not exhibit statistically significant results. This shows that using BraTS data alone does not suffice to generate a better performance for the model. Conversely, Experiment 1 and 9 shows that a good selection of out-of-domain data can bring about a positive increase in sensitivities. However, when comparing Experiment 6 (1308 patients) and Experiment 9 (481 patients), the Wilcoxon test found a significant difference, showing that a larger cohort of data does not necessarily guarantee better results. These results show that out-of-domain data must be curated carefully, as Experiment 6 contains multiple (i.e., 754 out of 910) cases that have the Not Categorized (NC) Breast Density category.

Therefore, for out-of-domain data to be useful, an appropriate selection of the breast density categories must be performed, as done in Experiment 9, which produces the best results.

6.5 Discussion

Our results indicate that short-range global information provides more contextual information to the network over its NLB counterpart for any configuration for the task of detecting metastases. The consistent improvements in FP rates indicate that the STROLL blocks provide “right-sized” self-attention to the network. While STROLL block demonstrates improvements in every configuration, placement at the earlier downsampling path is more advantageous as it allows more contextual information to be cascaded through the entire network. Additionally, STROLL blocks evaluate feature maps with reduced range, leading to lesser computational requirements, allowing the STROLL blocks to be placed at higher spatial resolutions compared to the NLB variant. We have also demonstrated that naive data fusion has positive effects on low volume datasets, despite the domain gap. The introduction of PHYSCISION data fusion strategy shows that supplementary data can provide important insights that lead to better curated datasets, which leads to leaner datasets and shorter training times. We observe that the D1 distribution, while giving us better sensitivity and FP, it does not constitute a perfect match to the SUNMET distributions, due to the morphology of primary gliomas which consists of larger volumes than brain metastases. However, cherry picking tumor diameters in BraTS is not possible, as a single 96x96x96 training patch contains multiple lesion sites, large and small. Reducing training patches to only capture a single lesion, results in smaller spatial views, which will be detrimental to training accuracy [99]. For the DUKE dataset, we found significant gains using BD over MT characteristics, as lesions in fatty breasts are easier to determine than dense ones [110]. In accordance with curriculum learning principles, our method also queues BD data points from most fatty to least fatty, so that samples increase in difficulty. MT did not fare as well, as it is a biological characteristic which has lesser influence on lesion appearance in MRI scans [111].

Similar work on brain metastases achieved 53% sensitivity at 8.3 FP/patient on a dataset with 121 scans trained on a multi-channel 2.5D GoogLeNet architecture [14]. Comparatively, STROLLNET a 3D scheme trained on single channel data achieves similar sensitivity at a lower FP rate of 2.56 FP/patient; we attribute this to the STROLL block which better discriminates smaller structures. A two-stage model trained on a significantly larger dataset with 3 MRI sequences obtained 90% sensitivity at 4.57 FP/patient [15]. To achieve comparable data volumes, the present study utilized a mono-channel single shot model to facilitate the input of different MRI modalities from different datasets. We favor single shot models over two-stage ones for simplicity of training and reduced computational complexity. Self-training schemes using brain MRI datasets achieved 74% sensitivity at 3.13 FP/patient after 3 iterations [112]. In contrast, PHYSCISION is trained once and achieves 2.13 FP/patient at a slightly reduced sensitivity, an acceptable tradeoff as the training time is reduced by a third.

A limitation of our study is the inability to harness the power of multi-channel deep learning as seen in [13], as a single channel implementation is more flexible in a multi-dataset application. Another limitation is the limited Breast Density information available in the DUKE dataset, which allowed only 17% of the data to be utilized. Future work would be explored should Breast Density information be available for more patients in this dataset.

6.6 Conclusion

To conclude, we have shown that modelling short range information is more beneficial than NLB in detecting brain metastases. In the opening statement, we highlighted the need for large volumes of data for deep learning. While this may be true for datasets with close domain gaps, it appears that for out-of-domain datasets, less is indeed more. Our work shows that curating the right data from disparate domains is crucial for efficient training of deep learning models, achieving performances comparable to that of iterative training on similar domain datasets.

7 Main Conclusion

The intricacies of lesion structures as well as the morphological differences between primary gliomas and secondary brain metastases necessitate the development of different strategies for the segmentation and detection of each lesion type. All three architectures investigated in this thesis share a commonality, the need for attention mechanisms to improve the performance of the task at hand. As much as possible, the attention blocks were designed to incorporate experts' soft knowledge to detect salient regions.

Also, the difficulty in obtaining fully annotated medical images hinders the curation of large pathological datasets to adequately train the model, particularly in single center studies. Therefore, every architecture presented in this thesis is developed together with a training paradigm to alleviate data scarcity. Auxiliary datasets from public medical image datasets were employed for this purpose. These datasets present varying domain gaps arising from different modalities, physiologies, organs, and even cancer primaries – providing an opportunity to explore domain adaptation techniques.

The CASPIAN block was derived from a-priori knowledge from the left-right asymmetry present in patients with primary gliomas. As lesions are irregularly shaped, asymmetry can be more pronounced in one or more reference planes. Multiplanar features were added to capture asymmetry on the 3 planes (coronal, axial, sagittal) in the CASPIANET++ model. This model improves segmentation of brain tumors over the winning BraTS 2018 model in WT and TC classes (81.08%, 91.20%, 87.60%, respectively for ET, WT and TC). The BraTS multiclass segmentation task is not immune to the issue of data imbalance and scarcity, particularly in the ET region. The Noisy Student Curriculum Learning (NSCL) was proposed to utilize the regularization effect of controlled data augmentation for training to further improve the model in the ET class (81.56%).

As CASPIANET++ model with NSCL was adept at brain segmentation to detect primary gliomas, we expanded the CASPIANET++ model to the SUNMET dataset. The SUNMET dataset consists of 121 patients containing secondary brain metastases (BM), with majority of cancers being of unknown primaries. This dataset is differentiated from others because it contains patients with metastatic brain tumors from a Malaysian demographic. The SUNMET brain MRIs are of a single MRI sequence compared to public datasets which are often multi-sequence. As there are differences in machine settings and requirements in medical practice, there always exists differences in MRI modalities used in each medical facility. Having a multi-channel deep learning architecture yields better results, but can be problematic when training multi-center, multi-organ and multi-pathology datasets. A single channel brain metastases detection method was favored for ease of transitioning to other datasets, which may not contain multimodal data. This tradeoff provides a more generalizable model, which potentially can be deployed to other medical centers.

Unfortunately, applying CASPIANET++ on SUNMET dataset yielded no tumor predictions. It was clear that a different architecture was required to detect secondary BM as it is too small to cause asymmetry in the brain. While the presence of primary gliomas can be described by a measure of symmetry in the brain, the same cannot be said about BM, the smallest being 2mm in size.

This led us to develop NLMET, which is based on N-Net, a detection architecture initially used for lung nodule detection. The original N-Net contains a location block that contains information of the lesion location as lung nodules are only found at the edge of the lungs. This pattern appears to be specific to lung nodules and not BMs; therefore non-local blocks were introduced to add more context to the model, to compensate for the redundant location block. Adding non-local blocks corresponded to the notion that BMs were present across a few slices in volumetric MRIs, which is out of scope of a local convolutional operation. The resulting architecture, NLMET registers better Sensitivity and FP/patient results over its N-Net counterpart (NLMET 64.9% Sensitivity at 3.304 FP/patient | N-Net 61.0% Sensitivity at 3.609 FP/patient). As FPs are a major stumbling block in the general adoption of CAD in the medical society, further improvements were achieved by incorporating gradual domain adaptation methods and through use of two volumetric brain MRI datasets. CAVEAT utilizes the BrainMetShare and BraTS datasets in a gradual self-training approach, which increases confidence scores in pseudo labels and increases the total volume of lesions seen by the model in every cycle of training, achieving a final Sensitivity of 74.0 at 3.131 FP/patient.

The non-local blocks in NLMET carrying information over larger ranges may inadvertently capture high inter Simensity regions such as blood vessels. Thus, STROLL blocks were introduced to evaluate non-local blocks at smaller ranges to capture salient regions more suited to smaller BM structures. The STROLLNET implementation achieves lower FP/patient than its non-local equivalent and in most configurations, gives at least 3% increase in sensitivity. While the NSCL and CAVEAT methods described previously require repetitive cycles - sometimes with unpredictable outcomes - a single pass data fusion method only requires one cycle and eliminates the uncertainty brought about by pseudo labels. PHYSCISION data fusion leverages on Big Data ideas, to pre-select quality datapoints through additional BiRADs breast cancer information and lesion diameter information in a single cycle of training to achieve 71.4 % Sensitivity at 2.130 FP/patient.

The three architectures presented in this thesis support the development of model agnostic blocks which can be implemented in other deep learning architectures, particularly in lesion segmentation and detection. The CASPIAN++ blocks are particularly suited for anatomical structures where symmetry is present. STROLL blocks are designed for detection of fine lesions in the brain and can be applied to detection of minute structures. It is also shown that medical data from various sources can be incorporated to boost the ability of deep learning systems. The NSCL and CAVEAT learning paradigms suggest that unlabeled medical data can be transformed into pseudo labels to further improve and generalize the model. The NSCL method is suited for datasets with vast amounts of unlabeled data, while CAVEAT utilizes a combination of 2 datasets with some domain gaps. While the former are iterative methods that require 4 cycles of training to obtain results, the PHYSCISION method incorporates ground truth data from disparate domains in a single pass data fusion approach. This is aided by metadata on attributes, which describe physical characteristics of the patient or its pathological structures.

7.1 Future Work

The work in this thesis can diverge into a few paths, across three major themes:

7.1.1 Architectural Improvements

CASPIANET++ describes a multi-channel deep learning architecture which uses four modalities on a single dataset. However, in the work proceeding it, we find that the multi sequence MRI information from different datasets cannot be reconciled for use in a multi-channel architecture as they may not contain the exact MRI modalities which are required. This resulted in the multi-modality MRI data being discarded and replaced with a single-channel deep learning architecture. The tradeoff for modality agnostic single-channel architectures is that they have better clinical applicability, but the information loss from multi-sequence MRI is apparent. Our attempts to utilize the random mix of modalities, to provide regularization to the model resulted in lower performance than using only the post-contrast MRI sequence, which fortunately can be found in each dataset. Previous work on missing modalities have been handled via dropouts [113] while in [100], single modalities were trained sequentially to achieve modality independent deep learning systems. Future work should extend into the development of adaptive channel architectures or utilizing skipped connections to cater for missing modalities. Another potential area is the development of GANs which can artificially enhance and normalize the missing inputs, this concept has been applied for data augmentation on scarcely annotated data [104].

7.1.2 Big Data Strategies, PyRadiomics and Morphological indicators

Our work on PHYSCISION explores the use of BiRADS data on the Duke Breast Cancer dataset to pre-select samples which will be beneficial to our training process. The next step would be to use full-fledged Big Data strategies, by incorporating more auxiliary metadata from patient records, especially when there are healthy/cancer patients in the dataset through unsupervised learning methods. While radiomics has traditionally been used for disease detection, the combination of radiomics and deep learning has been used with much success [114]. With the capability of an FCN to process large quantities of data as well as the availability of radiomics libraries such as PyRadiomics (https://www.radiomics.io/pyradiomics.html), this makes for an interesting area of study. The foundation of the CASPIANET++ block are concepts borrowed from classical CAD methods such as bilateral asymmetry. There are potentially more quantitative morphological indicators that can be obtained from the MRI scan to be used to develop architectural blocks.

7.1.3 Catastrophic Forgetting

Another open question of this study is to ascertain if there is an upper limit to the number of datasets that can be utilized before catastrophic forgetting occurs, or if the law of diminishing returns applies. Further domain adaptation work can be performed on previously identified (same organ, different pathology) ischemic stroke lesion (ISLES) and (different organ, different pathology) lung nodule (LUNA) datasets. Our work in NSCL and CAVEAT are semi-supervised learning methods and always feature brain MRI datasets in which data scarcity has always been an issue. Through transfer learning techniques, large medical datasets of different organs and pathologies can be used to learn high level representations while a second refinement stage is implemented specifically for brain MRI datasets. Work in [61] explores this extensively for various datasets, and produces exceptional results on lung nodule detection, but did not fare as well for the brain tumor datasets.

Bibliography

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” *CA. Cancer J. Clin.*, vol. 72, no. 1, pp. 7–33, Jan. 2022.

[2] G. P. Mazzara, R. P. Velthuizen, J. L. Pearlman, H. M. Greenberg, and H. Wagner, “Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 59, no. 1, pp. 300–312, May 2004.

[3] D. E. Morris, J. D. Bourland, J. G. Rosenman, and E. G. Shaw, “Three-dimensional conformal radiation treatment planning and delivery for low- and intermediate-grade gliomas,” *Semin. Radiat. Oncol.*, vol. 11, no. 2, pp. 124–137, 2001.

[4] T. Jeremy Crawford *et al.*, “Expertise Affects Inter-Observer Agreement at Peripheral Locations within a Brain Tumor,” *Front. Psychol*, vol. 8, p. 1628, 2017.

[5] Y. Wu, W. Yang, J. Jiang, S. Li, Q. Feng, and W. Chen, “Semi-automatic segmentation of brain tumors using population and individual information,” *J. Digit. Imaging*, vol. 26, no. 4, pp. 786–796, Aug. 2013.

[6] K. Kamnitsas *et al.*, “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[7] L. Nayak, E. Q. Lee, and P. Y. Wen, “Epidemiology of brain metastases,” *Curr. Oncol. Rep.*, vol. 14, no. 1, pp. 48–54, Feb. 2012.

[8] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, 27-May-2015.

[9] L. Strohm, C. Hehakaya, E. R. Ranschaert, W. P. C. Boon, and E. H. M. Moors, “Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors,” *Eur. Radiol.*, vol. 30, no. 10, pp. 5525–5532, Oct. 2020.

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.”

[11] M. Havaei *et al.*, “Brain tumor segmentation with Deep Neural Networks,” *Med. Image Anal.*, vol. 35, pp. 18–31, May 2017.

[12] M. Zhang *et al.*, “Deep-Learning Detection of Cancer Metastases to the Brain on MRI,” *J. Magn. Reson. Imaging*, vol. 52, no. 4, pp. 1227–1236, Oct. 2020.

[13] S. T. Jünger *et al.*, “Fully Automated <scp>MR</scp> Detection and Segmentation of Brain Metastases in Non-small Cell Lung Cancer Using Deep Learning,” *J. Magn. Reson. Imaging*, p. jmri.27741, May 2021.

[14] E. Grøvik, D. Yi, M. Iv, E. Tong, D. Rubin, and G. Zaharchuk, “Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI,” *J. Magn. Reson. Imaging*, vol. 51, no. 1, pp. 175–182, Jan. 2020.

[15] R. Li *et al.*, “MRI-based two-stage deep learning model for automatic detection and segmentation of brain metastases,” *Eur. Radiol. 2023*, pp. 1–11, Jan. 2023.

[16] E. Dikici *et al.*, “Automated Brain Metastases Detection Framework for T1-Weighted Contrast-Enhanced 3D MRI,” *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 10, pp. 2883–2893, Oct. 2020.

[17] S. Hussain, S. M. Anwar, and M. Majid, “Segmentation of glioma tumors in brain using deep convolutional neural network,” *Neurocomputing*, vol. 282, pp. 248–261, Mar. 2018.

[18] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images,” in *Advances in Neural Information Processing Systems*, 2012, vol. 4, pp. 2843–2851.

[19] M. Seyedhosseini, M. Sajjadi, and T. Tasdizen, “Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks,” in *Proceedings of the IEEE International Conference on*

Computer Vision, 2013, pp. 2168–2175.

[20] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 3431–3440.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,” 2014.

[22] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, pp. 234–241.

[23] T. M. Quan, D. G. C. Hildebrand, and W.-K. Jeong, “FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics,” Dec. 2016.

[24] H. Chen, X. Qi, J. Z. Cheng, and P. A. Heng, “Deep contextual networks for neuronal structure segmentation,” in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 1167–1173.

[25] A. Myronenko, “3D MRI brain tumor segmentation using autoencoder regularization,” 2018.

[26] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, “No New-Net,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11384 LNCS, pp. 234–244, Sep. 2018.

[27] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[28] K. Kamnitsas *et al.*, “Ensembles of multiple models and architectures for robust brain tumour segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10670 LNCS, pp. 450–462.

[29] M. Rezaei, H. Yang, and C. Meinel, “Voxel-GAN: Adversarial framework for learning imbalanced brain tumor segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11384 LNCS, pp. 321–333.

[30] M. Rezaei, H. Yang, and C. Meinel, “Multi-Task Generative Adversarial Network for Handling Imbalanced Clinical Data,” 2018.

[31] P. Moeskops *et al.*, “Deep Learning for Multi-Task Medical Image Segmentation in Multiple Modalities,” Apr. 2017.

[32] H. R. Roth *et al.*, “A New 2.5D Representation for Lymph Node Detection using Random Sets of Deep Convolutional Neural Network Observations,” Jun. 2014.

[33] C. Zhou, C. Ding, Z. Lu, X. Wang, and D. Tao, “One-pass multi-task convolutional neural networks for efficient brain tumor segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11072 LNCS, pp. 637–645.

[34] J. Xue *et al.*, “Deep learning-based detection and segmentation-assisted management of brain metastases,” *Neuro. Oncol.*, vol. 22, no. 4, pp. 505–514, Apr. 2020.

[35] S. Grand, C. Pasteris, A. Attye, J. F. Le Bas, and A. Krainik, “The different faces of central nervous system metastases,” *Diagn. Interv. Imaging*, vol. 95, no. 10, pp. 917–931, Oct. 2014.

[36] E. Dikici *et al.*, “Automated Brain Metastases Detection Framework for T1-Weighted Contrast-Enhanced 3D MRI,” *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, Aug. 2019.

[37] Z. Zhou *et al.*, “Computer-aided Detection of Brain Metastases in T1-weighted MRI for Stereotactic Radiosurgery Using Deep Learning Single-Shot Detectors,” *Radiology*, p. 191479, Mar. 2020.

[38] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, “Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network.”

[39] O. Charron, A. Lallement, D. Jarnet, V. Noblet, J. B. Clavier, and P. Meyer, “Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network,” *Comput. Biol. Med.*, vol. 95, pp. 43–54, Apr. 2018.

[40] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-December, pp. 5999–6009.

[41] J. Schlemper *et al.*, “Attention gated networks: Learning to leverage salient regions in medical images,” *Med. Image Anal.*, vol. 53, pp. 197–207, Aug. 2019.

[42] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[43] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” Apr. 2018.

[44] A. G. Roy, N. Navab, and C. Wachinger, “Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11070 LNCS, pp. 421–429, Mar. 2018.

[45] N. J. Tustison *et al.*, “Optimal Symmetric Multimodal Templates and Concatenated Random Forests for Supervised Brain Tumor Segmentation (Simplified) with ANTsR,” *Neuroinformatics*, vol. 13, no. 2, pp. 209–225, Nov. 2015.

[46] M. Erihov, S. Alpert, P. Kisilev, and S. Hashoul, “A cross saliency approach to asymmetry-based tumor detection,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, pp. 636–643.

[47] H. Chen, Z. Qin, Y. Ding, L. Tian, and Z. Qin, “Brain tumor segmentation with deep convolutional symmetric neural network,” *Neurocomputing*, vol. 392, pp. 305–313, Jun. 2020.

[48] N. Ray, B. N. Saha, and M. R. G. Brown, “Locating brain tumors from MR imagery using symmetry,” in *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 2007, pp. 224–228.

[49] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, “One-pass Multi-task Networks with Cross-task Guided Attention for Brain Tumor Segmentation,” *IEEE Trans. Image Process.*, vol. 29, pp. 4516–4529, Jun. 2019.

[50] A. Tao, N. Karan, S. Nvidia, and B. Catanzaro Nvidia, “HIERARCHICAL MULTI-SCALE ATTENTION FOR SEMANTIC SEGMENTATION,” 2020.

[51] A. Sinha and J. Dolz, “Multi-scale self-guided attention for medical image segmentation,” *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, Jun. 2019.

[52] X. Wang, R. Girshick, A. Gupta, and K. He, *Non-local Neural Networks*. IEEE Computer Society, 2018, pp. 7794–7803.

[53] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[54] D.-H. Lee, “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks,” 2013.

[55] W. Bai *et al.*, “Semi-supervised learning for network-based cardiac MR image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10434 LNCS, pp. 253–260.

[56] Y. Zhou *et al.*, “Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training,” in *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, 2019, pp. 121–140.

[57] E. Arazo, D. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning,” Aug. 2019.

[58] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Label Propagation for Deep Semi-supervised Learning,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 5065–5074, Apr. 2019.

[59] M. Huh, P. Agrawal, and A. A. Efros, “What makes ImageNet good for transfer learning?”

[60] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” *Adv. Neural Inf. Process. Syst.*, vol. 32, Feb. 2019.

[61] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, “Models Genesis,” *Med. Image Anal.*, vol. 67, 2021.

[62] B. Zoph *et al.*, “Rethinking Pre-training and Self-training,” Jun. 2020.

[63] A. Liew, C. C. Lee, B. L. Lan, and M. Tan, “CASPIANET++: A multidimensional Channel-Spatial Asymmetric attention network with Noisy Student Curriculum Learning paradigm for brain tumor segmentation,” *Comput. Biol. Med.*, vol. 136, p. 104690, Sep. 2021.

[64] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, “Self-Loop Uncertainty: A Novel Pseudo-Label for Semi-Supervised Medical Image Segmentation,” 2020.

[65] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, “DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8661–8670, Aug. 2021.

[66] J. Wacker, M. Ladeira, and J. Eduardo Vaz Nascimento, “Transfer Learning for Brain Tumor Segmentation.”

[67] D.-H. Lee, “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.”

[68] T. Bdair, N. Navab, and S. Albarqouni, “ROAM: Random Layer Mixup for Semi-Supervised Learning in Medical Imaging.”

[69] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-Training With Noisy Student Improves ImageNet Classification,” 2020, pp. 10684–10695.

[70] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” 2009.

[71] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, “Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11046 LNCS, pp. 249–258, Jul. 2018.

[72] J. De Fauw *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nat. Med. 2018 249*, vol. 24, no. 9, pp. 1342–1350, Aug. 2018.

[73] Y. Zhu *et al.*, “Improving Semantic Segmentation via Self-Training,” Apr. 2020.

[74] L. Yang, L. Liu, Y. Wang, P. Wang, and Y. Zhang, “Multi-Domain Joint Training for Person Re-Identification,” Jan. 2022.

[75] E. C. Holland, “Progenitor cells and glioma formation,” *Current Opinion in Neurology*, vol. 14, no. 6. pp. 683–688, 2001.

[76] H. Ohgaki and P. Kleihues, “Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas,” *Journal of Neuropathology and Experimental Neurology*, vol. 64, no. 6. Lippincott Williams and Wilkins, pp. 479–489, 2005.

[77] S. Bakas *et al.*, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” Nov. 2018.

[78] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci. Data*, vol. 4, Sep. 2017.

[79] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans.*

Med. Imaging, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[80] M. Lin, K. Tang, and X. Yao, “Dynamic sampling approach to training neural networks for multiclass imbalance classification,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 24, no. 4, pp. 647–660, 2013.

[81] Y. Feng, M. Zhou, and X. Tong, “Imbalanced classification: an objective-oriented review,” Feb. 2020.

[82] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, Nov. 2013.

[83] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” 2002.

[84] H. Small and J. Ventura, “Handling Unbalanced Data in Deep Image Segmentation,” 2017.

[85] R. Guerrero *et al.*, “White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks,” *NeuroImage Clin.*, vol. 17, pp. 918–934, Jun. 2018.

[86] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, and X. Lladó, “Acute and sub-acute stroke lesion segmentation from multimodal MRI,” 2018.

[87] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10553 LNCS, pp. 240–248.

[88] L. Fidon *et al.*, “Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation using Holistic Convolutional Networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10670 LNCS, pp. 64–76, Jul. 2017.

[89] U. Baid *et al.*, “The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification,” Jul. 2021.

[90] C. Chen, X. Liu, M. Ding, J. Zheng, and J. Li, “3D Dilated Multi-Fiber Network for Real-time Brain Tumor Segmentation in MRI.”

[91] Y. Qin *et al.*, “Autofocus Layer for Semantic Segmentation.”

[92] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, Oct. 2017.

[93] G. Ghiasi *et al.*, “Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation.”

[94] F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, “nnU-Net for Brain Tumor Segmentation,” Nov. 2020.

[95] R. McKinley, R. Meier, and R. Wiest, “Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11384 LNCS, pp. 456–465.

[96] Y. Wang *et al.*, “Modality-Pairing Learning for Brain Tumor Segmentation,” Springer, Cham, 2021, pp. 230–240.

[97] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015.

[98] R. G. Ellenbogen, S. I. Abdulrauf, and L. N. Sekhar, “Principles of neurological surgery,” *Princ. Neurol. Surg.*, pp. 1–820, Apr. 2012.

[99] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,” *Biomed. Opt. Express*, vol. 9, no. 7, p. 3049, Jul. 2018.

[100] S. Thakur *et al.*, “Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training,” *Neuroimage*, vol. 220, p. 117081, Oct. 2020.

[101] A. A. A. Setio *et al.*, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge,” *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.

[102] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond,” *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 1971–1980, Apr. 2019.

[103] S. Mittal, S. Galesso, and T. Brox, “Essentials for Class Incremental Learning.”

[104] C. Han *et al.*, “Learning More with Less: Conditional PGGAN-based Data Augmentation for Brain Metastases Detection Using Highly-Rough Annotation on MR Images,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 119–127, Feb. 2019.

[105] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci. data*, vol. 4, Sep. 2017.

[106] A. Saha *et al.*, “A machine learning approach to radiogenomics of breastcancer: a study of 922 subjects and 529 DCE-MRI features,” *Br. J. Cancer*, vol. 119, no. 4, p. 508, Aug. 2018.

[107] J. Witowski *et al.*, “Improving breast cancer diagnostics with deep learning for MRI,” *Sci. Transl. Med.*, vol. 14, no. 664, Sep. 2022.

[108] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local Neural Networks,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 7794–7803, Nov. 2017.

[109] D. O. Kamson *et al.*, “Differentiation of glioblastomas from metastatic brain tumors by tryptophan uptake and kinetic analysis: A positron emission tomographic study with magnetic resonance imaging comparison,” *Mol. Imaging*, vol. 12, no. 5, pp. 327–337, Jul. 2013.

[110] S. S. Nazari and P. Mukherjee, “An overview of mammographic density and its association with breast cancer,” *Breast Cancer*, vol. 25, no. 3, pp. 259–267, May 2018.

[111] O. Yersal and S. Barutca, “Biological subtypes of breast cancer: Prognostic and therapeutic implications,” *World J. Clin. Oncol.*, vol. 5, no. 3, p. 412, Aug. 2014.

[112] A. Liew, C. C. Lee, V. Subramaniam, B. L. Lan, and M. Tan, “Gradual Self-Training via Confidence and Volume Based Domain Adaptation for Multi Dataset Deep Learning-Based Brain Metastases Detection Using Nonlocal Networks on MRI Images,” *J. Magn. Reson. Imaging*, 2022.

[113] E. Grøvik *et al.*, “Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study,” *npj Digit. Med. 2021 41*, vol. 4, no. 1, pp. 1–7, Feb. 2021.

[114] H. H. Yeoh *et al.*, “RADIFUSION: A multi-radiomics deep learning based breast cancer risk prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement,” Apr. 2023.