

White Matter Multiple Sclerosis Lesion Segmentation Under Distributional Shifts



By

Ali Haider

Registration # 00000327532

Supervisor

Dr. Syed Omer Gilani

Department of Biomedical Engineering & Sciences
School of Mechanical and Manufacturing Engineering (SMME)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

August 2023

White Matter Multiple Sclerosis Lesion Segmentation Under Distributional Shifts



By

Ali Haider

Registration # 00000327532

Supervisor

Dr. Syed Omer Gilani

Co-supervisor

Dr. Asim Waris

Co-supervisor

Dr. Adeeb Shehzad

A thesis submitted in conformity with the requirements for
the degree of *Master of Science* in
Biomedical Engineering

Department of Biomedical Engineering & Sciences
School of Mechanical and Manufacturing Engineering (SMME)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

August 2023

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by **Regn No. 00000327532 Ali Haider** of **School of Mechanical & Manufacturing Engineering (SMME) (SMME)** has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis titled. **White Matter Multiple Sclerosis (MS) lesion segmentation in 3D Magnetic Resonance Imaging (MRI) of the brain with Distributional Shift**


Signature:  -

Name (Supervisor): Syed Omer Gilani

Date: 11 - Aug - 2023

Signature (HOD):  -

Date: 11 - Aug - 2023

Signature (DEAN):  -

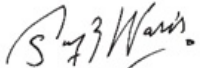
Date: 11 - Aug - 2023





National University of Sciences & Technology (NUST)
MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: Ali Haider (00000327532)
Titled: White Matter Multiple Sclerosis (MS) lesion segmentation in 3D Magnetic Resonance Imaging (MRI) of the brain with Distributional Shift be accepted in partial fulfillment of the requirements for the award of MS in Biomedical Engineering degree.

Examination Committee Members

1. Name: Muhammad Asim Waris Signature: 

2. Name: Adeeb Shehzad Signature: 

Supervisor: Syed Omer Gilani
Signature: 
Date: 11 - Aug - 2023



Head of Department

11 - Aug - 2023

Date

COUNTERSIGNED

11 - Aug - 2023

Date



Dean/Principal

.....

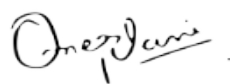
Proposed Certificate for Plagiarism

It is certified that MS Thesis Titled White Matter Multiple Sclerosis Lesion Segmentation Under Distributional Shifts by Ali Haider has been examined by us. We undertake the follows:

- a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
- b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
- c. There is no fabrication of data or results which have been compiled/analyzed.
- d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Syed Omer Gilani

Signature :  _____

Date: 18-Aug-2023

Declaration

I, *Ali Haider* assert that this thesis titled "White Matter Multiple Sclerosis Lesion Segmentation Under Distributional Shifts" and the content within it stems from my original investigations and are solely the product of my own efforts and research.

I confirm that:

1. This research was primarily undertaken during my candidacy for a Master of Science degree at NUST.
2. Wherever a portion of this dissertation has been presented earlier for a degree or any other credentials at NUST or any different institution, it has been distinctly indicated.
3. Whenever I've referenced the published contributions of others, I've always distinctly acknowledged them.
4. Whenever I've cited directly from others' works, I've always provided the source. Apart from these direct quotes, this dissertation is wholly the product of my own efforts.
5. I have recognized and given credit to all primary sources of assistance.
6. In instances where the dissertation involves collaborative work, I've explicitly detailed the contributions made by others and outlined my personal input.

Ali Haider

Ali Haider

00000327532

Copyright Notice

- The copyright for the content of this thesis belongs to the student author. Reproductions, whether in entirety or in parts, can only be executed based on the directions provided by the author and stored in the SMME Library at NUST. Information on this can be acquired from the Librarian. This page should be included in any such reproductions. Any further reproductions cannot be undertaken without the explicit written consent of the author.
- The intellectual property rights mentioned in this thesis belong to SMME, NUST, unless there's a previous agreement stating otherwise. Third parties cannot utilize these rights without the express written consent of SMME. Any agreement will outline the specific terms and conditions of use.
- Additional details regarding the circumstances for revealing and utilizing the content can be obtained from the SMME Library at NUST, Islamabad.

This thesis is dedicated to *my beloved parents*

Abstract

In the rapidly evolving era of machine learning and deep learning, new algorithms are constantly emerging, each built upon existing research and pushing the boundaries in the field of medical imaging. However, one of the major challenges in the application of these algorithms is the distributional shifts that occur in real-world datasets. This research paper utilizes the expanded Shifts 2.0 dataset that was released for The Shift Challenge 2022. It presents how to enhance the UNET model's robustness and uncertainty estimations in the segmentation of white matter lesions in Multiple Sclerosis patients, using only the FLAIR modality. This approach examines the impact of multiple hyperparameters on the results of the Shift 2.0 dataset. The suggested model yielded R-AUC scores of 1.12 and 1.60 on the Dev-out and Eval-out of the shift dataset, in contrast to the baseline UNET method which registered scores of 4.66 and 7.40 on those respective partitions. Moreover, the paper establishes that the performance of an ensemble of UNET models can be comparable to that of a transformer-based ensemble of UNETR models, offering promising implications for future research and applications.

Keywords: *Multiple Sclerosis, Semantic Segmentation, Distributional Shift, UNET*

Acknowledgments

We would like to express our deepest gratitude to The Shifts Project for their remarkable efforts in organizing The Shifts Challenge 2022 and The University of Lausanne (UNIL) for the evaluation dataset. This challenge has provided researchers from all over the world with a unique opportunity to engage and practice real-life distributional shifts. We are immensely grateful to the organizers for their dedication, vision, and commitment in creating this exceptional platform for knowledge exchange and progress.

Contents

1	Introduction	1
1.1	Overview	1
1.1.1	Multiple Sclerosis	2
1.1.2	Magnetic Resonance Imaging in Multiple Sclerosis	5
1.2	Distributional Shifts	6
2	Methodology & Results	8
2.1	Dataset	9
2.1.1	Pre-processing	9
2.2	Baseline architecture	10
2.3	Proposed architecture	10
2.4	Implementation	11
2.5	Results	12

CONTENTS

2.5.1	Evaluation and Matrices	12
2.5.2	Hyperparameter Tuning	12
2.6	Discussion	13
2.7	Conclusions	16
	References	18

List of Figures

1.1	A sample FLAIR image that shows the Multiple Sclerosis White Matter lesions and their corresponding ground truth labels shown in red	2
1.2	Picture taken from National Multiple Sclerosis Society website	3
1.3	Different types of Multiple sclerosis.	4
1.4	Difference between synthetic and Real Distributional shifts (Image from shifts project)	7
2.1	Difference between synthetic and Real Distributional shifts (Image from shifts project)	11
2.2	Comparison of the baseline and the proposed architecture’s R-AUC results on different shifts dataset partitions. Proposed Eval_out is compared with baseline dev_out as the baseline Eval_out values were not available	13
2.3	Comparison of the baseline and the proposed architecture’s nDSC results on different shifts dataset partitions. Proposed Eval_out is compared with baseline dev_out as the baseline Eval_out values were not available	14
2.4	Comparison of the baseline and the proposed architecture’s F1-score results on different shifts dataset partitions	14

List of Tables

2.1	Shows the canonical distribution and the number of patients in each partition of the Shifts dataset.	9
2.2	Demonstrating the effect of different numbers of samples on the robustness and uncertainty R-AUC(%).	13
2.3	Robustness and Uncertainty R-AUC(%) of the ensembles of three models. Showing the comparison of baseline 32 samples with 128 samples and different sizes.	13
2.4	Most of the Tabel is taken from the original paper [1]. Segmentation effectiveness (nDSC) and combined assessment of robustness and uncertainty (R-AUC) of baseline and proposed frameworks. Ensembles are made from the combination of 5 models in all cases except the Evl_out for the proposed architecture an ensemble of only 3 models is used.	17

List of Abbreviations and Symbols

Abbreviations

MRI	Magnetic Resonance Imaging
FLAIR	Fluid Attenuated Inversion Recovery
MS	Multiple Sclerosis
WML	White matter lesions
ML/DL	Machine Learning/Deep Learning
DUA	Data usage agreement
MONAI	Medical Open Network for Artificial Intelligence
nDSC	Normalised dice similarity coefficients
R-AUC	Area under the retention curve

Introduction

1.1 Overview

It's often assumed in the realm of machine learning that training, validation, and test datasets are separate and evenly distributed. This implies that strong test results are a vital marker of the model's efficacy in deployment. The distributional shift is the incongruity between training and real-life dataset. It is one of the most common issues faced by data scientists while implementing machine learning algorithms. Unknown or uncontrollable factors cause most distributional shifts. Machine learning models should ideally exhibit robust generalization across various distributional shifts. However, because of the no-free lunch theorem [2], it is impractical to resist all shifts.

Multiple Sclerosis (MS) is a persistent, untreatable, and advancing ailment of the central nervous system that profoundly affects a person's well-being and daily living. It is an auto-immune disorder in which myelin is attacked resulting in scar tissue known as sclerosis and hindering the nerve's ability to translate impulses. The tissues of interest are present as white in the FLAIR images, which can be seen in Fig 1. In the same figure, the bottom row provides a visual representation of the truth labels, depicted in red and superimposed onto the original image. MS lesion delineation pertains to the process of creating a 3D pixel-by-pixel segmentation mask for brain abnormalities, using either single or multimodal MRI images.

Occasionally, medical institutions share patient information, and training pictures are scarce for machine-learning purposes. No openly accessible dataset wholly captures the disease’s variation in terms of intensity and progression, restricting the effectiveness and resilience of machine learning frameworks in real-world scenarios. Moreover, alterations in MRI machine manufacturers, magnetic field setups, or imaging programs can lead to differences in MRI images in terms of pixel dimensions, signal-to-noise ratio, contrast settings, slice thickness, non-linearity corrections, and so on. Variations in the image-gathering and labeling procedure can also be exacerbated by changes in the clinicians operating the devices. These disparities get worsened when the images acquired from several medical centers are combined, constituting a substantial distributional shift for ML-based algorithms.

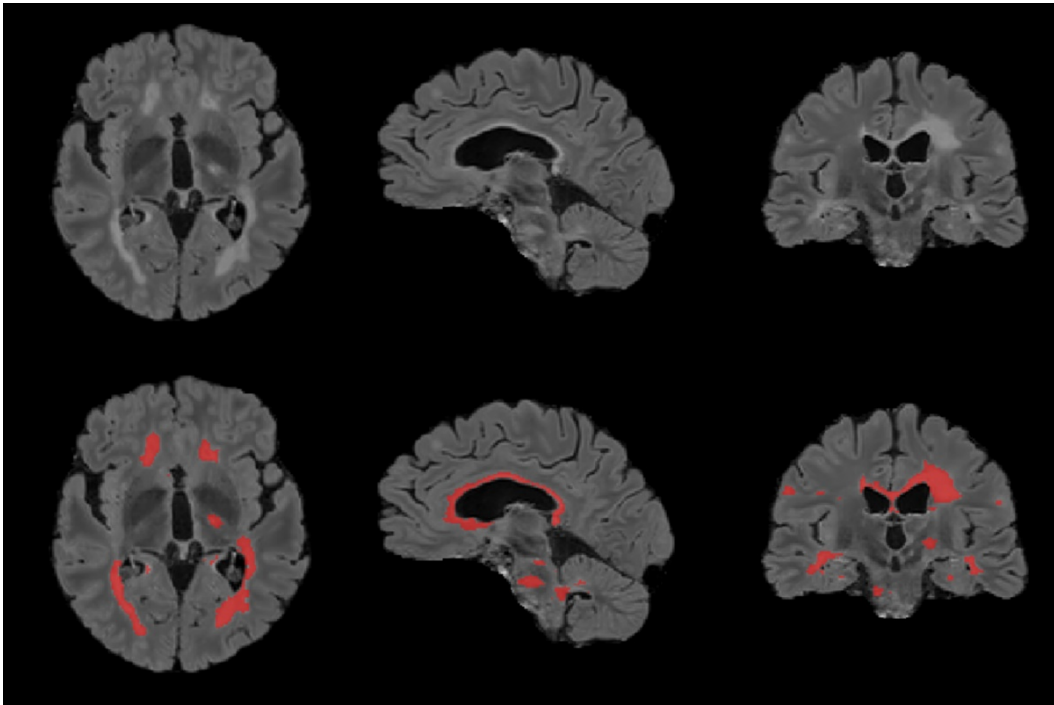


Figure 1.1: A sample FLAIR image that shows the Multiple Sclerosis White Matter lesions and their corresponding ground truth labels shown in red

1.1.1 Multiple Sclerosis

Multiple sclerosis impacts the central nervous system, encompassing the brain, spinal cord, and optic pathways. While the precise origin of MS remains elusive, some factor prompts the immune system to target the CNS. The breakdown of myelin, a protec-

tive sheath around nerve strands, disrupts communication to and from the brain. This interruption of signal transmission results in an array of symptoms, such as sensations of numbness, tingling, mood fluctuations, memory challenges, pain, fatigue, vision loss, and/or paralysis. Each individual's experience with MS is distinct, and these impairments can either be transient or lasting.

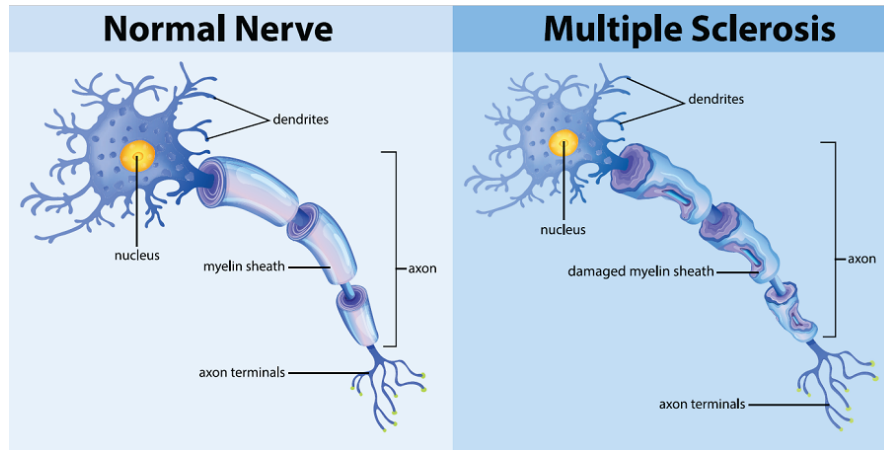


Figure 1.2: Picture taken from National Multiple Sclerosis Society website

Multiple sclerosis is an unpredictable ailment of the central nervous system that hinders the transmission of data within the brain and between the brain and the body. As established by the International Advisory Committee on Clinical Trials of MS in 1996, it may be classified into four categories or disease courses. They are as follows:

- Clinically isolated syndrome (CIS)
- Relapsing-remitting MS
- Secondary progressive MS
- Primary progressive MS

A clinically isolated syndrome (CIS) is the first incidence of neurologic symptoms produced by inflammation and demyelination in the central nervous system. CIS symptoms differ from one to person, however, it typically includes, Eyesight and vision problems, sensation loss in the face, arm and leg weakness, coordination and balance loss, and problems with bladder control.

Relapsing-remitting MS (RRMS) is the most common form of MS, characterized by clear episodes of new or worsening neurological symptoms, known as relapses, followed by recovery periods or remissions where symptoms can either fade or become permanent. The disease doesn't progress during remissions. RRMS can be active, with relapses or new MRI activity, or inactive; it can also be deteriorating post-relapse or stable. Initially, around 85% of MS patients are diagnosed with RRMS, which can later progress to a secondary stage.

Secondary progressive MS (SPMS) follows relapsing-remitting MS (RRMS), where some individuals experience a gradual decline in neurological function or accumulating impairments. SPMS can be active, marked by relapses or new MRI findings, or inactive, and can either show the progression of disability over time, with or without relapses and MRI changes or remain without such progression.

From the onset of symptoms in Primary Progressive MS (PPMS), neurological function declines without initial relapses or remissions. PPMS can be active, showcasing occasional relapses or new MRI activity, or inactive, and might exhibit continuous disability progression, regardless of relapses or MRI changes. Approximately 15% of MS patients are affected by PPMS.

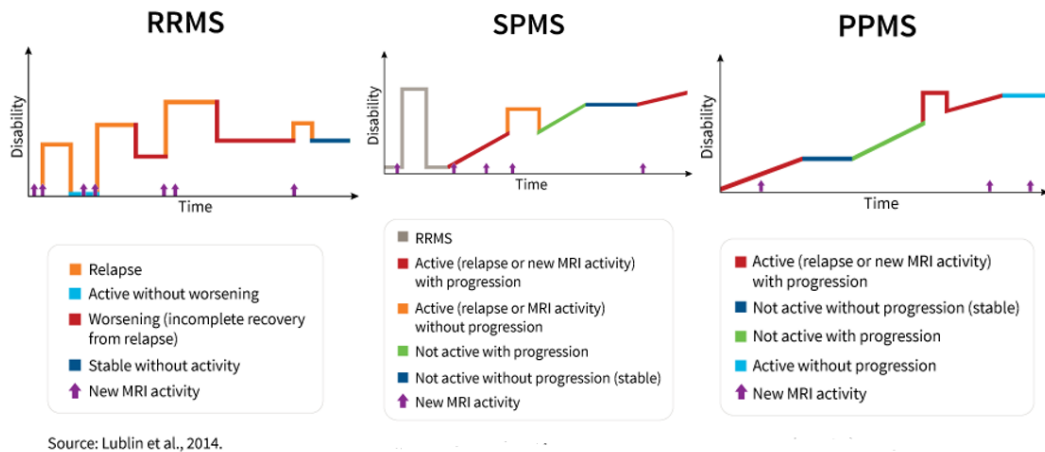


Figure 1.3: Different types of Multiple sclerosis.

1.1.2 Magnetic Resonance Imaging in Multiple Sclerosis

Magnetic resonance imaging (MRI) is currently the most advanced, non-invasive tool for visualizing the brain, spinal cord, and other body regions. It is chiefly used to diagnose MS and track its development. Through MRI, we've gained deeper insights into the effects of MS and expanded our understanding of the disease.

In contrast to computed tomography (CT) scans or standard X-rays, MRI doesn't rely on radiation. Instead, it uses magnetic fields and radio waves to detect the amount of water in the body's tissues. Since the myelin sheath safeguarding nerve fibers contains fat and repels water, areas damaged by MS, where this fat layer is lost, retain more water. These regions appear as either bright white or dark spots on an MRI image, based on the scanning method applied.

More specifically, the MRI functions as follows:

1. An intense magnetic field makes a minor portion of the hydrogen protons in water molecules align with the field's direction.
2. After alignment, radio waves and supplementary weaker magnetic fields disturb this alignment.
3. Once these waves cease, the protons revert to their original alignment. During this reversion, they emit signals that a computer interprets to produce an image.

Multiple MRI scan varieties are employed for MS. On occasions, gadolinium, a contrasting medium, is infused into the bloodstream during the MRI process to pinpoint newly inflamed regions. Since gadolinium consists of large molecules, it typically can't breach the blood-brain barrier, which acts as a shield preventing materials from transitioning from the blood to the central nervous system. But if there's active inflammation, this barrier can be compromised, allowing gadolinium to seep through and emphasize the inflamed spots.

Typical MRI methods utilized in MS encompass:

1. T-1 weighted without gadolinium — might reveal darkened spots (hypointensities) which possibly denote permanent nerve injuries.
2. T-1 weighted with gadolinium — could display luminous spots (enhancing lesions) pointing to ongoing inflammation.
3. presents the comprehensive disease impact or the entirety of lesions, both ancient and recent.
4. Fluid attenuated inversion recovery (FLAIR) — depicts MS dynamics by diminishing spinal fluid-related disruptions.

1.2 Distributional Shifts

Distributional shift refers to the disparity between data used for training and that during actual use. Adapting to 'shifted' data can be tough for machine learning models. As the extent of this shift grows, the efficacy of ML models usually declines. For instance, both AI and human drivers taught to operate vehicles on the right might face challenges in territories where driving is on the left. Recognizing distributional shifts is prevalent in machine learning and is particularly crucial in applications where safety is paramount.

predictability uncertainties have centered on compact image classification datasets. These datasets often exhibit synthetic or non-authentic types of distributional changes. Findings from such datasets infrequently translate to large-scale ML implementations. The absence of expansive, varied, and industry-derived datasets showcasing genuine distributional changes hinders the validation of novel methods and the derivation of insights that are relevant to practical applications.

Optimally, machine learning models ought to adapt effectively across various distributional changes. If they don't, these models should signal this through uncertainty measurements, helping us implement strategies to enhance the system's safety and dependability. Methods promoting consistent or equivalent representations, averting feature loss, and enabling more comprehensive data insights can boost adaptability.

Approaches producing metrics responsive to the magnitude of distributional alterations will provide more accurate uncertainty assessments.



Figure 1.4: Difference between synthetic and Real Distributional shifts (Image from shifts project)

Methodology & Results

3D MRI segmentation of MS involves the delineation of white matter lesions from the surrounding healthy tissue within the volumetric MRI data. This process aids in assessing the disease’s progression and the patient’s response to treatment. The original paper [1] introduced a foundational approach that relies on a 3D UNET framework [3] and transformer-based architecture of the UNETR [4]. UNET, a seminal model in medical image segmentation, functions via two distinct paths: an encoder, which progressively reduces the resolution, and a decoder, which restores the resolution while simultaneously reducing the number of features. This results in semantic segmentation. Several critical parameters influence the output of a U-Net model, including the depth of convolutional layers, strides, and the use of residual blocks, among others. On the other hand, UNETR capitalizes on the power of the transformer as an encoder to capture global scale information, incorporating skip connections to link the encoder and decoder at varying resolutions, while also preserving the U-Net’s trademark U-shaped architecture. The hyperparameters of the baseline method were tuned in accordance with [5]. A deep ensemble [6] is created involving the aggregation of output probabilities from five distinct UNet and UNETR models. The baseline result shows that the performance of the transformer-based model is much better than the U-Net model as exhibited in Table 2.4.

2.1 Dataset

The Shifts Project MS segmentation dataset integrates numerous open-access datasets under a unified data usage agreement (DUA). Included in these databases are PubMRI [7], ISBI [8, 9], and MSSEG-1 [10]. However, the dataset supplied by the University of Lausanne wasn't made public because of patient security considerations. Instead, it was utilized for assessment via Docker in the Shifts Challenge. Typically, machine learning datasets are divided into training, evaluation, and test segments.

However, in this case, the dataset includes training, development-in, development-out, evaluation-in, and evaluation-out sections. The 'in' and 'out' suffixes denote whether the dataset falls within the domain or outside it (shifted). Regardless of whether the subject belongs to an in-domain or out-domain dataset, each one includes two modalities: FLAIR and T1w. Additionally, some datasets also incorporate T1w contrast-enhanced, proton density, and T2w modalities. For training, both the baseline method [11] and the method proposed in this research utilized exclusively the FLAIR modality, adhering to the consensus recommendation [12]. An overview of the distribution and the number of patients in each partition of the Shifts dataset is outlined in Table 2.1.

Type	In-Domain			Out-Domain	
Data	Train	Dev_in	Eval_in	Dev_out	Eval_out
Patients	33	7	33	25	74

Table 2.1: Shows the canonical distribution and the number of patients in each partition of the Shifts dataset.

2.1.1 Pre-processing

The Shifts dataset, a composite of diverse sub-datasets featuring distinct resolutions, scanning devices, and magnetic strengths, necessitated thorough preprocessing to standardize MRI images. This includes denoising [13], skull stripping [14], and bias field correction [15]. To assist with precise segmentation, brain masks were generated by registering the T1 modality image to the FLAIR space [16]. Finally, all images underwent interpolation to establish a uniform 1mm isovoxel space, ensuring consistency across the dataset and enabling accurate analysis.

2.2 Baseline architecture

Data augmentation is a technique employed in deep learning systems to introduce minor modifications to the training data, helping prevent overfitting and enhancing results on testing datasets. The image augmentation in this case is done by MONAI [17], in which multiple transforms are applied to the training dataset. In this process, 32 samples of size $[96 \times 96 \times 96]$ voxels are cropped from each training image with the center being the lesion voxel. The primary structure of the U-Net model is a symmetrical architecture composed of five sets of encoding and decoding layers, each associated with progressively escalating channel dimensions starting from 32, doubling to 64, then 128, 256, and finally peaking at 512. The model incorporates a strategy of stride-based downsampling, wherein strides of 2 are uniformly applied in every dimension, aiming to incrementally reduce the spatial dimensions while augmenting the depth of feature maps. The model is designed with zero residual units, avoiding the potential complexity and computational load brought by them.

The total trainable parameters in the model are 7,912,874. Overall 5 separate instances of the models were trained, each for a total of 300 epochs. The results are tabulated and detailed in Table 2.4, providing a comparative view of the performance metrics across the different models and configurations like UNET, Monte Carlo dropouts, and UNETR in Single form and in the ensemble of 5 models.

2.3 Proposed architecture

The performance of the proposed U-Net retains the symmetric architecture characteristic of UNet models, with five encoding and decoding layers corresponding to the growing channel dimensions: 32, 64, 128, 256, and 512. A significant change lies in the stride-based downsampling strategy, where the initial layer adopts a stride of 1, followed by strides of 2 for the subsequent layers, resulting in a more gradual reduction in spatial dimensions. The model includes two residual units, which contrasts with the previous model that had none, adding an additional level of complexity and changing the number

of trainable parameters to 19,216,097.

The input size to this model has been reduced to a spatial dimension of $48 \times 48 \times 48$, which is a significant decrease compared to the previous model. This model is trained for a duration of 20 to 30 epochs under a dynamic learning rate and early stopping. When compared with the previous model, it is similar in overall structure but has notable differences in the stride in the first encoding layer and the inclusion of residual units. This potentially impacts on model performance, with the ability to capture more local information due to the slower initial downsampling, and enhanced feature extraction capacity from the inclusion of residual units. This model’s computational efficiency might be higher due to reduced input size, but it is also reduced by the inclusion of residual blocks. The training time per epoch is around 430 seconds. As an ensemble of 5 models is used therefore the total time for training is around 15 hours. The batch size is the same for both baseline and proposed systems. The block diagram of the proposed model is shown in Figure 2.1.

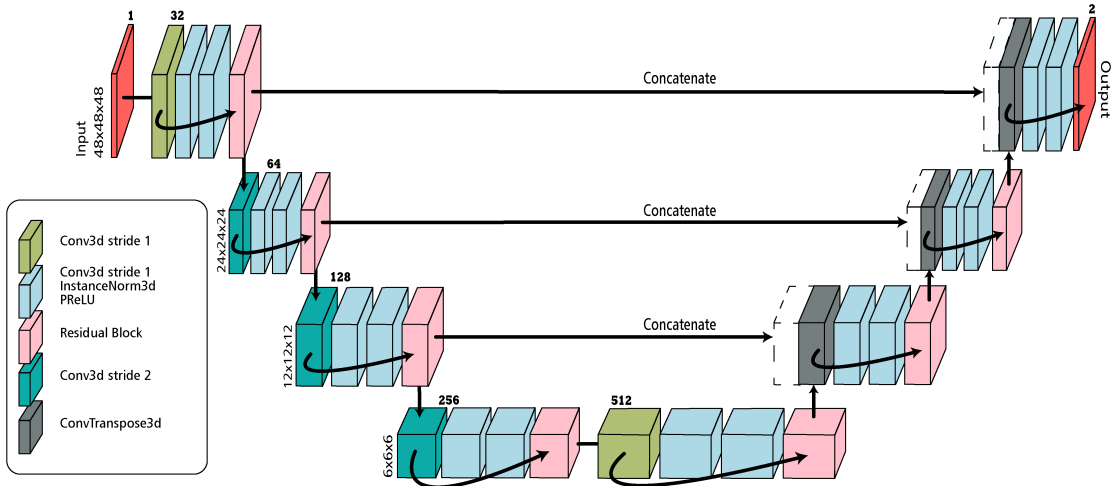


Figure 2.1: Difference between synthetic and Real Distributional shifts (Image from shifts project)

2.4 Implementation

Hardware used for training includes NVIDIA GTX 1080 Ti 12GB GPU, Intel E5520 @ 2.27 GHz x 8 processor, and 16 GB RAM. The coding language used is Python language in the MONAI framework [17] based on Pytorch [18]). The code requires Python version

3.9 along with CUDA and other libraries

2.5 Results

2.5.1 Evaluation and Matrices

Usually, the 3D MRI image segmentation is evaluated by the dice similarity coefficient [19, 20]. But one of the major drawbacks of DSC is that its value depends on the size of the lesion load. So baseline method proposed an adaptive (normalized) Dice similarity coefficient (nDSC) which decouples the predictive accuracy from the size of the lesion. Error retention curves [6, 21, 22] are used to examine both robustness and uncertainty. These are error metrics that represent the error in decreasing order of uncertainty when the predictions of a model are substituted by the ground truth labels. The model’s predictive performance is determined by the area under the error retention curves, which can be reduced by either enhancing the model’s ability to predict, resulting in lower overall error, or by presenting better uncertainty estimates associated with an error. Hence, the area below the error retention curves (R-AUC) is an indicator that measures both resistance to the distributional shift and the quality of uncertainty.

$$DSC = \frac{2|Y_1 \cap Y_2|}{|Y_1| + |Y_2|} = \frac{2TP}{FP + 2TP + FN} = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (2.5.1)$$

$$\overline{Pr}_{\tau^*} = \frac{TP_{\tau^*}}{TP_{\tau^*} + k_p FP_{\tau^*}} \quad (2.5.2)$$

2.5.2 Hyperparameter Tuning

Shifts dataset only uses 33 patient’s data for training purposes which for deep learning algorithms is still small. To cater to this issue a MONAI transform is used which crops multiple samples of specified size from a single image. During tuning it is discovered

that the best number of samples is around 128 per image and the size of each sample is around 48x48x48 as shown in Table 2.2 and Table 2.3.

Sample Size	Number of samples	R-AUC (%)
48x48x48	64	1.5845 ± 0.8336
48x48x48	128	1.3098 ± 1.0751
48x48x48	160	1.3725 ± 0.8597

Table 2.2: Demonstrating the effect of different numbers of samples on the robustness and uncertainty R-AUC(%).

Number of samples	Size of sample	R-AUC (%)
32	96x96x96	2.9190 ± 1.7687
128	32x32x32	1.4274 ± 0.8569
128	48x48x48	1.3098 ± 1.0751
128	64x64x64	1.3147 ± 0.8311
128	72x72x72	1.4385 ± 0.9586

Table 2.3: Robustness and Uncertainty R-AUC(%) of the ensembles of three models. Showing the comparison of baseline 32 samples with 128 samples and different sizes.

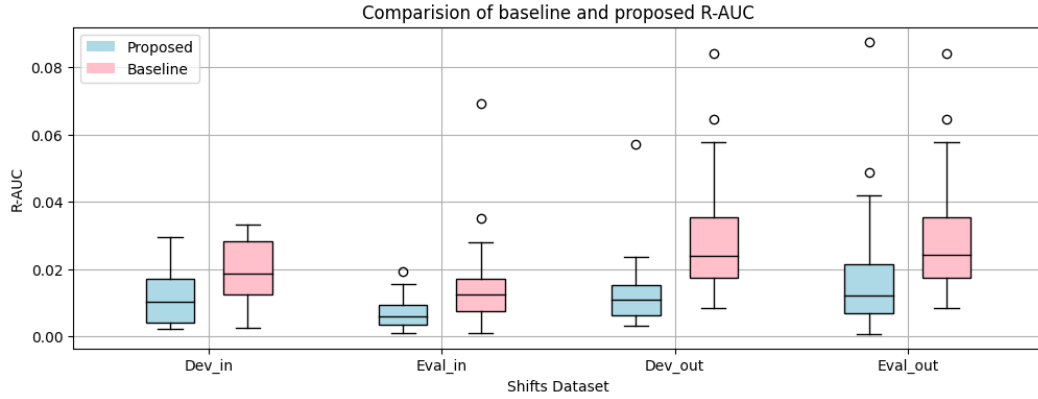


Figure 2.2: Comparison of the baseline and the proposed architecture’s R-AUC results on different shifts dataset partitions. Proposed Eval_out is compared with baseline dev_out as the baseline Eval_out values were not available

2.6 Discussion

Recent advancements in deep learning have considerably simplified the diagnosis and prognosis of Multiple Sclerosis (MS). Like many machine learning paradigms, deep learning techniques for MS are prone to distributional shifts due to disparities between training and test datasets. Although many deep learning algorithms show commendable generalization across diverse datasets, the baseline method employed uncertainty estimations to delineate the model’s performance over specific datasets.

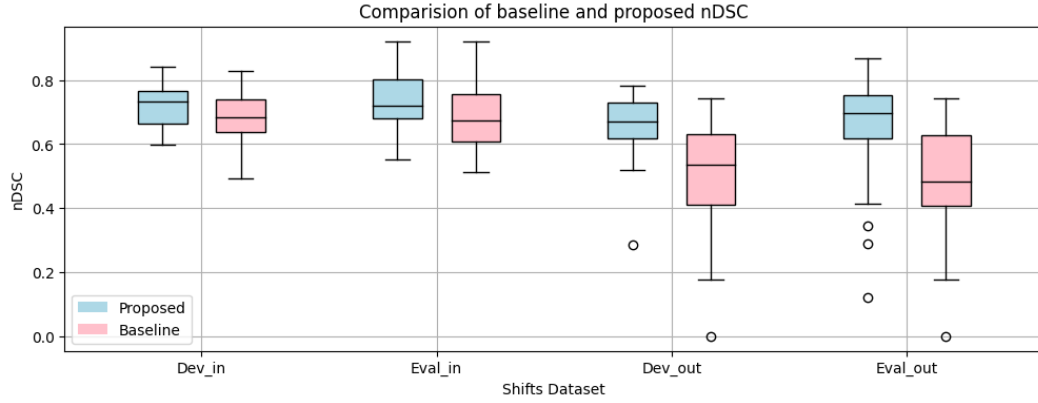


Figure 2.3: Comparison of the baseline and the proposed architecture’s nDSC results on different shifts dataset partitions. Proposed Eval_out is compared with baseline dev_out as the baseline Eval_out values were not available

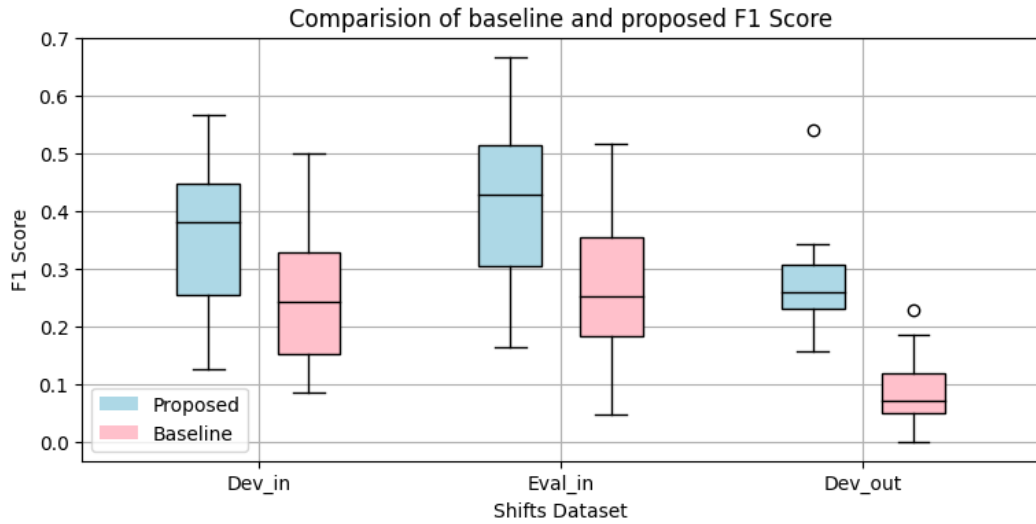


Figure 2.4: Comparison of the baseline and the proposed architecture’s F1-score results on different shifts dataset partitions

The Shifts dataset, which this study utilized, possesses distinctive partitioning: it’s segmented into in-domain and out-domain datasets. The in-domain segment comprises the conventional train, evaluation, and test divisions. In contrast, the out-domain is bifurcated into two test datasets, with one publicly available and the other designated for online evaluations via Docker. Essential preprocessing steps, such as denoising, skull stripping, and isovoxelation, were applied to the data.

In Baseline, [1] an ensemble of UNet and UNETR models was used. The findings, as presented in Table 2.4, underscore the superior efficacy of the transformer-based model (UNETR) over the conventional UNet model across all Shifts dataset partitions. The

baseline UNet configuration incorporated 5 layers with channels 32, 64, 128, 256, and 512. Downsampling with a stride of 2 was executed in the encoder phase. Notably, no residual units were integrated. The model utilized 32 samples derived from a single input image of dimensions 96x96x96.

Conversely, our proposed model mirrors the U-shaped architecture of UNet but with some notable modifications. While maintaining the same channel count, we integrated residual blocks and adjusted the stride for the initial layer to 1. Dropout remained unaltered at zero. Additionally, our model harnessed 128 samples extracted from an input image sized 48x48x48. Both dynamic learning rate adjustments and early stopping mechanisms (at a Dice similarity coefficient of 0.70) were implemented. Tables 2.2 and 2.3 further delve into the adjustments concerning sample size and number. A similar sort of behavior can be seen from residual blocks in [23]

The cumulative results, presented in Table 2.4’s bottom row, spotlight the superior performance of our model’s ensemble over the baseline UNet ensemble. When compared with the UNETR ensemble, our method exhibited enhanced performance in every partition of the dataset, with the exception of dev-in.

In alignment with consensus recommendations, our study solely utilized the FLAIR modality for training. The potential of dual-modal training, incorporating both FLAIR and T1W images, remained unexplored due to system constraints, particularly RAM and GPU memory limitations. Combining various architectural approaches might lead to enhanced outcomes[24]. Nonetheless, our research highlights the potential avenues and implications this methodology can open up for future studies and real-world applications.

In this research paper, we present our participation in the Shifts Challenge 2022 conducted by The Shifts Project. We opted for task 2 of the challenge where our proposed model achieved an impressive 5th and 8th position on the leaderboard. The task was to develop a deep-learning system that can demonstrate its robustness and also predicts its uncertainty on a real-world problem like the segmentation of white matter lesions in Multiple sclerosis. Furthermore, our findings highlight the importance of feature engineering as a critical aspect of model development enabling us to extract maximum

performance from the given dataset

2.7 Conclusions

In this research paper, we present our participation in the Shifts Challenge 2022 conducted by The Shifts Project. We opted for task 2 of the challenge where our proposed model achieved an impressive 5th and 8th position on the leaderboard. The task was to develop a deep-learning system that can demonstrate its robustness and also predicts its uncertainty on a real-world problem like the segmentation of white matter lesions in Multiple sclerosis. Furthermore, our findings highlight the importance of feature engineering as a critical aspect of model development enabling us to extract maximum performance from the given dataset. The outcomes substantiate the effectiveness of our proposed methods, affirming their potential to achieve performances that are comparable to the existing baseline UNETR model. The proposed architecture is not only accurate in segmenting the desired features but also provides reliable and consistent predictions across different datasets.

Type	Model	$nDSC(\%) \uparrow$				$R - AUC(\%) \downarrow$			
		Dev_in	Dev_out	Evl_in	Evl_out	Dev_in	Dev_out	Evl_in	Evl_out
Single	UNET	68.54	49.33	67.59	55.79	2.51	7.84	2.77	9.87
	UNET-DP	59.73	48.35	63.93	54.43	2.62	8.76	2.66	9.71
	UNETR	71.21	51.60	69.27	56.76	1.89	6.17	1.95	6.47
Ensemble	UNET	69.70	50.85	68.89	57.53	1.17	4.66	1.76	7.40
	UNET-DP	60.65	44.70	61.78	50.06	1.92	6.77	2.52	7.89
	UNETR	72.51	53.46	71.41	59.49	0.34	1.52	0.63	2.88
Proposed	UNET	71.75	66.05	73.25	67.67	1.16	1.12	0.62	1.60

Table 2.4: Most of the Tabel is taken from the original paper [1]. Segmentation effectiveness (nDSC) and combined assessment of robustness and uncertainty (R-AUC) of baseline and proposed frameworks. Ensembles are made from the combination of 5 models in all cases except the Evl_out for the proposed architecture an ensemble of only 3 models is used.

References

- [1] Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark J. F. Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, Nataliia Molchanova, Antonis Nikitakis, Vatsal Raina, Francesco La Rosa, Eli Sivena, Vasileios Tsarsitalidis, Efi Tsompopoulou, and Elena Volf. Shifts 2.0: Extending The Dataset of Real Distributional Shifts. June 2022. doi: 10.48550/arXiv.2206.15407.
- [2] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020.
- [3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer International Publishing. ISBN 978-3-319-46723-8.
- [4] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, March 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210310504H>. Accepted to IEEE Winter Conference on Applications of Computer Vision (WACV) 2022.
- [5] F. La Rosa, A. Abdulkadir, M. J. Fartaria, R. Rahmanzadeh, P. J. Lu, R. Galbusera, M. Barakovic, J. P. Thiran, C. Granziera, and M. B. Cuadra. Multiple sclerosis cortical and wm lesion segmentation at 3t mri: a deep learning method based on flair and mp2rage. *Neuroimage Clin*, 27:102335, 2020. ISSN 2213-1582. doi: 10.1016/j.nicl.2020.102335.
- [6] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and

- scalable predictive uncertainty estimation using deep ensembles, December 01, 2016 2016. URL <https://ui.adsabs.harvard.edu/abs/2016arXiv161201474L>. NIPS 2017.
- [7] Ž Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, and Ž Špičin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. 2018. ISSN 1539-2791. 1559-0089 *Neuroinformatics*. 2018 Jan;16(1):51-63. doi: 10.1007/s12021-017-9348-7.
- [8] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C. A. M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Ihome, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P. L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage*, 148:77–102, 2017. ISSN 1053-8119 (Print) 1053-8119. doi: 10.1016/j.neuroimage.2016.12.064.
- [9] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Pierre-Louis Bazin, Peter A. Calabresi, Ciprian M. Crainiceanu, Lotta M. Ellingsen, Daniel S. Reich, Jerry L. Prince, and Dzung L. Pham. Longitudinal multiple sclerosis lesion segmentation data resource. *Data in Brief*, 12:346–350, 2017. ISSN 2352-3409. doi: <https://doi.org/10.1016/j.dib.2017.04.004>.
- [10] Olivier Commowick, Audrey Istace, Michaël Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Améli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Tristan Glatard, Jérémy Beaumont, Senan Doyle, Florence Forbes, Jesse Knight, April Khademi, Amirreza Mahbod, Chunliang Wang, Richard McKinley, Franca Wagner, John Muschelli, Elizabeth Sweeney, Eloy Roura, Xavier Lladó, Michel M. Santos, Wellington P. Santos, Abel G. Silva-Filho, Xavier Tomas-Fernandez, Hélène Urien, Isabelle Bloch, Sergi Valverde, Mariano Cabezas, Francisco Javier Vera-Olmos,

- Norberto Malpica, Charles Guttman, Sandra Vukusic, Gilles Edan, Michel Dojat, Martin Styner, Simon K. Warfield, François Cotton, and Christian Barillot. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):13650, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-31911-7.
- [11] Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark J. F. Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, Nataliia Molchanova, Antonis Nikitakis, Vatsal Raina, Francesco La Rosa, Eli Sivena, Vasileios Tsarsitalidis, Efi Tsompopoulou, and Elena Volf. Shifts 2.0: Extending The Dataset of Real Distributional Shifts. June 2022. doi: 10.48550/arXiv.2206.15407.
- [12] M. P. Wattjes, O. Ciccarelli, D. S. Reich, B. Banwell, N. de Stefano, C. Enzinger, F. Fazekas, M. Filippi, J. Frederiksen, C. Gasperini, Y. Hachohen, L. Kappos, D. K. B. Li, K. Mankad, X. Montalban, S. D. Newsome, J. Oh, J. Palace, M. A. Rocca, J. Sastre-Garriga, M. Tintoré, A. Traboulsee, H. Vrenken, T. Yousry, F. Barkhof, and À Rovira. 2021 magnims-cmsc-naims consensus recommendations on the use of mri in patients with multiple sclerosis. *Lancet Neurol*, 20(8):653–670, 2021. ISSN 1474-4422. doi: 10.1016/s1474-4422(21)00095-8.
- [13] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. 27 (4):425–41, 2008. 1558-254x IEEE Trans Med Imaging. 2008 Apr;27(4):425-41. doi: 10.1109/TMI.2007.906087.
- [14] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H. P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kickingreder. Automated brain extraction of multisequence mri using artificial neural networks. 1097-0193 Hum Brain Mapp. 2019 Dec 1;40(17):4952-4964. doi: 10.1002/hbm.24750. Epub 2019 Aug 12.
- [15] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: improved n3 bias correction. *IEEE Trans Med Imaging*. 2010 Jun;29(6):1310-20. doi: 10.1109/TMI.2010.2046908. Epub 2010 Apr 8.
- [16] O. Commowick, N. Wiest-Daesslé, and S. Prima. Block-matching strategies for

- rigid registration of multimodal medical images. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 700–703. ISBN 1945-8452. doi: 10.1109/ISBI.2012.6235644.
- [17] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Flock, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. Monai: An open-source framework for deep learning in healthcare, November 01, 2022 2022. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv221102701C>. www.monai.io.
- [18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, December 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191201703P>. 12 pages, 3 figures, NeurIPS 2019.
- [19] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. ISSN 00129658, 19399170. doi: 10.2307/1932409. URL <http://www.jstor.org/stable/1932409>.
- [20] Tage Sørensen, Tage Sørensen, Tor Biering-Sørensen, Tia Sørensen, and John T. Sorensen. A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on danish commons. 1948.

- [21] Andrey Malinin. Uncertainty estimation in deep learning with application to spoken language assessment. 2019.
- [22] Andrey Malinin, Neil Band, Ganshin, Alexander, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Roginskiy, Denis, Mariya Shmatova, Panos Tigas, and Boris Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks, July 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210707455M>.
- [23] K. Zafar, S. O. Gilani, A. Waris, A. Ahmed, M. Jamil, M. N. Khan, and A. Sohail Kashif. Skin lesion segmentation from dermoscopic images using convolutional neural network. *Sensors (Basel)*, 20(6), 2020. ISSN 1424-8220. doi: 10.3390/s20061601. *Sensors (Basel)*. 2020 Mar 13;20(6):1601.
- [24] Mahnoor Ali, Syed Omer Gilani, Asim Waris, Kashan Zafar, and Mohsin Jamil. Brain tumour image segmentation using deep networks. *IEEE Access*, 8:153589–153598, 2020. doi: 10.1109/ACCESS.2020.3018160.