Leveraging Metadata in Ischaemic Stroke MRI Segmentation Using FiLM Conditioning Layers

Ilakya Prabhakar

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

Stroke cases have reached epidemic levels, with 1 in 4 adults over the age of 25 standing to have one within their lifetime. Accurately segmented lesions imaged with MRI not only assist in making early intervention decisions, but are vital for establishing accurate long term prognosis and rehabilitation plans. Automatic segmentation algorithms can both scale to create large neuroimaging datasets which aid rehabilitation research, and on an individual level they ensure precise outcomes where time is of the essence. Image segmentation algorithms tend to be unimodal in nature [1, 2], only considering the image as an input. To investigate the effect of incorporating multi-modal information, this dissertation uses the ATLAS dataset [3], as it provides annotated ischaemic stroke lesions alongside patient and image metadata (such as stroke laterality, and days post stroke MRI was taken). This research explores the effects of conditioning a baseline U-Net [1] model with these different types of metadata, using a technique called Feature-wise Linear Modulation (FiLM) [4] to modulate image features in the network with tabular patient and image metadata. The aim is to investigate what types of metadata are useful to condition a network with, and if there are specific cases where conditioning with metadata helps more than others. It is found that incorporating a mix of available and derived (from ground truth) metadata results in an increase in Dice score of 7.3% over baseline. Even though FiLM layers cannot directly encode spatial information, spatial metadata such as Stroke Location and Stroke Laterality prove the most effective available metadata to condition the network with, implying that lesions have distinct visual features in different anatomical regions. Conditioning networks with metadata is also found to have the most improvement over the baseline for cases where lesions are almost entirely missed.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Ilakya Prabhakar*)

Acknowledgements

I would like to extend my thanks to Canon Medical Research Europe for proposing and co-supervising this dissertation research. Specifically, my supervisor Dr. Alison O'Neil for her valuable time spent offering guidance and constructive criticism, and Antanas Kascenas for his many hours spent offering solutions and help with implementation issues. I would also like to thank my internal co-super supervisor Dr. Henry Gouk for academic advice, and offering a non-medical research perspective. I am very fortunate to have acquired such a diverse range of knowledge and skills from these individuals during the course of this dissertation. Finally, I would like to express my gratitude to Google DeepMind, whose financial support has allowed me to pursue this Master's program.

Table of Contents

Chapter 1

Introduction

Worldwide, stroke stands to be the leading cause of adult disability, and second leading cause of death [5], with ischaemic stroke [1] specifically comprising 87% of all cases. With up to two thirds of all strokes leading to severe disability, early interventions in the acute stage along with accurate prognosis at the post-acute stage are vital for long term recovery and rehabilitation. Medical images, whether MRI or CT, play a crucial part, especially with a growing body of research establishing connections between anatomical lesion location and eventual stroke outcome [6]. This knowledge can be used to guide further therapeutic decisions, and assess potential for long-term recovery of motor/speech capabilities. Lesion delineation on a medical image (a task herein referred to as 'segmentation') is usually manually performed by a radiologist, however, the process is painstakingly slow and the quality varies greatly depending on the neuroimaging experience of the annotator. Some patients can end up waiting 24 hours to receive a clinical image analysis, which in the acute stages is a critical amount of time. Thus, the motivation for automatic and accurate lesion segmentation is clear - automatic segmentation can both scale across large neuroimaging datasets to aid rehabilitation research, and on an individual patient level it ensures fast and accurate clinical outcomes in time critical cases.

State of the art medical image segmentation models use deep learning algorithms based on convolutional neural networks [1, 2] due to their ability to learn hierarchical and spatially invariant image features. These models typically rely on smaller datasets with preprocessing methods that apply extensive data augmentation, so as to extract all the available information in each image [2]. However, often image and patient *metadata* is

[1] Ischaemic stroke, is a type of stroke that occurs when there is a sudden blockage or narrowing of an artery supplying blood to the brain. The alternative type of stroke is called a 'hemorrhagic stroke' which occurs when there is bleeding in or around the brain.

neglected as an information source when performing automatic segmentation. Metadata can include information pertaining to the MRI itself, such as the scanner type, or can also be patient specific information such as how long after stroke onset the MRI has been acquired. It is logical that this supplementary information would aid in the automatic segmentation task - supported by the findings of Boonn et al. [7] who report that radiologists themselves greatly benefit from information such as patient history, but are hindered by the time it takes to acquire this. This dissertation will therefore investigate the efficacy of leveraging this metadata in ischaemic stroke lesion segmentation by using a multi-modal deep learning model.

Multi-modal [2] approaches to image segmentation have been attempted in other fields. In the task of object detection in autonomous vehicles, Person et al. [8] combine classic image segmentation with a CNN with LiDAR point cloud data to create a fusion model with decision trees. In the task of video classification by content, Trzcinski finds that incorporating both visual and textual features into the model results in a model improvement from the baseline single-modality model of up to 95% [9]. The motivation for exploiting all the available information in a segmentation task is clear. A successful approach to incorporating multi-modal information in a deep learning task is introduced by Perez et al. [4] by using Feature-wise Linear Modulation (FiLM) layers. These layers allow for feature-wise enhancement/suppression, conditioned on an arbitrary input. These layers apply an affine transformation to the intermediate features in the CNN, and the parameters of this transformation are defined by the output of a network which processes the conditioning information - the image and patient metadata in this case. [3]

The dataset used in this dissertation will be the publicly available ATLAS (Anatomical Tracings of Lesions After Stroke) dataset [3]. The ATLAS dataset is a manually segmented set of T1-weighted MRI images, created in order to promote research for automated lesion segmentation of brain MRI scans. This dataset is chosen due to the accompanying range of diverse metadata that is provided, including aspects such as chronicity (number of days post stroke that the MRI was acquired), laterality (side of stroke), anatomical region and image acquisition parameters. This research will integrate FiLM conditioning layers into a state-of-the-art biomedical segmentation deep learning model, the U-Net [1], in order to investigate the effect of incorporating the

[2] In this dissertation we take multi-modal to refer to combining image with tabular numerical or textual data, as opposed to combining different modalities of imaging.

[3] This paragraph is taken verbatim from Section 1 of the IPP report.

metadata present in the ATLAS dataset on lesion segmentation accuracy.

There are two strands of investigation. The first concerns the metadata itself. The hypothesis is that certain types of metadata might be more effective to include than others, with others potentially having a negative impact. Thus, an evaluation will be conducted to establish the quantitative and qualitative effect of including each metadata type on resulting segmentation predictions.

- What kinds of metadata help when input into a segmentation model, and why?
- Are there particular lesions for which the additional metadata benefits the automatic segmentation model more so than others, and what are their characteristics?

Part of this controlled evaluation will be to derive metadata from the annotated ground truth itself, to establish an upper bound on how useful injected metadata can be in improving predictions. Examples of derived metadata include average lesion pixel intensity, lesion volume, or lesion location. These quantities can be roughly mapped to real clinical data such as chronicity (related to lesion pixel intensity) and symptom locations (related to lesion location/size). The second strand of investigations concerns the architectural integration of the FiLM layers themselves. The conditioning layers can be placed on only some, or all layers of the U-Net. It is hypothesised that the conditioning layer placement will also have an effect on model performance. Thus the core research question is: What effect does both the type of metadata, and layer positioning of FiLM layers have on model performance during stroke lesion segmentation?

The paper is structured as such: in the second chapter we introduce the task of medical image segmentation and describe at a high level the concepts involved. Multimodal approaches to image segmentation are then discussed, leading into a mathematical explanation of the specific approach of using FiLM conditioning layers. Chapter 3 focuses on the methodology of the research undertaken. First, the dataset is introduced with some preliminary analysis conducted into the metadata. Synthetic metadata is then derived from ground truth and lesion characteristics, with clinical justifications provided for each derived value. The model architecture, training scheme, and parameter choices are next explained in detail, and the chapter ends with a description of how results are evaluated. Chapter 4 presents results from all conducted experiments with accompanying theoretical analysis, and the final chapter gives the conclusions of the dissertation along with suggestions for future research directions.

Chapter 2

Background

2.1 Ischaemic Stroke Imaging

Ischaemic stroke occurs when there is a disruption of blood flow to a specific part of the brain, caused by a blockage or narrowing of a blood vessel, often due to a blood clot or a buildup of fatty deposits in the arteries that supply blood to the brain. The lack of blood flow results in a lack of oxygen and nutrients and in turn damages brain cells in the affected area. To view the extent of this damage and plan intervention and rehabilitation strategies, both Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are used. In this thesis, MRI images will be used, specifically T1-Weighted MRI. On T1-weighted MRI images, ischaemic strokes typically appear as areas of low signal intensity, which means they appear darker compared to surrounding normal brain tissue. The terms used to describe axes within the brain are shown in *Fig. 2.1*, and will be used throughout the dissertation. MRI images are typically oriented in RAS+ space, where the Right, Anterior and Superior directions are the positive axial directions in 3D space.

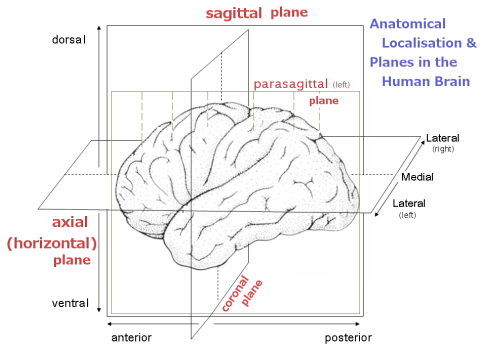


Figure 2.1: Conventional neurological terms used to describe different planes and directional axes within the brain.

2.2 Biomedical Image Segmentation

In the field of computer vision, image segmentation refers to the process of dividing a digital image into multiple segments, in order to transform it into something more representative and easier to analyse, depending on the task at hand. In the biomedical context, this usually involves detecting the boundaries of anatomical structures, lesions, or organs for diagnostic or therapeutic purposes. Segmentation datasets are comprised of pairs of image and 'ground truth' (*Fig. 2.2*) wherein the ground truth is a pixel-level label array of identical size to the image in question and the label value indicates which structure or lesion the associated pixel belongs to. In a binary segmentation task there is only one foreground label of interest. For medical images, the ground truth is usually costly to obtain, requiring clinical experts to painstakingly trace regions of the 3D image both slice-by-slice, and pixel-by-pixel. Thus, automated methods for biomedical image segmentation usually rely on small datasets and methods of extracting all possible salient features from these.

(a) Original MRI (b) Ground truth (c) MRI with overlay

Figure 2.2: Image and ground truth pair taken from the ATLAS dataset [3].

The most popular deep learning model for biomedical image segmentation is the U-Net [1], named as such for its U-shaped autoencoder architecture and popularised due to its effectiveness and simplicity to implement. Most state-of-the-art improvements have been variations of this architecture [2][10].

The autoencoder design, shown in *Fig. 2.3*, comprises a contracting path that compresses image representations into a latent space of much lower dimensionality, and a decoder path which consists of transposed convolutional layers to upsample the latent representation. The input is a 2D or 3D image, and output is a segmentation map. The almost symmetric expanding path allows for a high resolution pixelwise output. Skip connections between the contracting and expanding paths are a key component of the

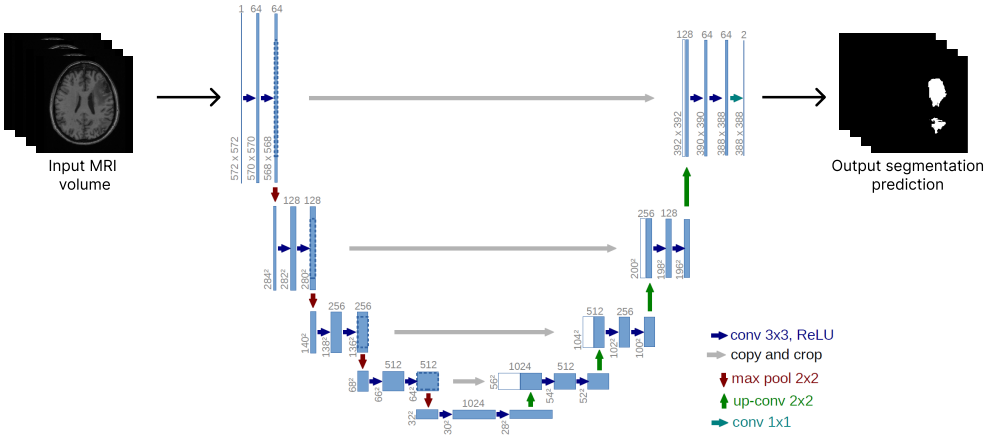


Figure 2.3: Diagram depicting U-Net architecture, taken from [1].

U-net, and ensure that the output segmentation can be a combination of both higher resolution features, and also very deep latent feature representations. This allows the network to accurately segment, by bypassing the need to fully decode exact pixels and lines from a compressed latent representation. The standard loss function in the U-Net for a segmentation task is computed by first applying a pixel-wise softmax to the output predictions, and then comparing these pixel-wise to the ground truth using the cross entropy loss function (Eqn 3.1). This essentially treats the segmentation problem as a pixel-wise classification problem. [1]

2.3 Multi-Modal Inputs to Segmentation Networks

When radiologists use medical images to make clinical decisions, these are often contextualised with patient specific information from electronic health records (EHRs) before coming to a diagnosis [7]. Despite this, within deep learning image segmentation research there is a bias towards unimodal architectures which only use the images themselves as inputs. This is partly due to the lack of publicly available datasets which include patient and demographic information - datasets usually have a requirement of keeping patients anonymous, and also there is a lot of work required to curate this additional metadata which often comes from different sources to the image itself. However, it is also due to the lack of definitively successful performance when incorporating this information into existing unimodal architectures.

[1] This paragraph is taken from the Section 2.2 of the IPP.

2.3.1 Feature Concatenation

Previous work using multi-modal inputs has used the technique of feature concatenation - shown in *Fig. 2.4*. This is a method of combining representations from different modalities without having to train separate models. Features are extracted separately for each modality, and concatenated or stacked before entering the classifier or decoder stage of the network, thus the ability to use the same target label is retained, even with multiple input types.

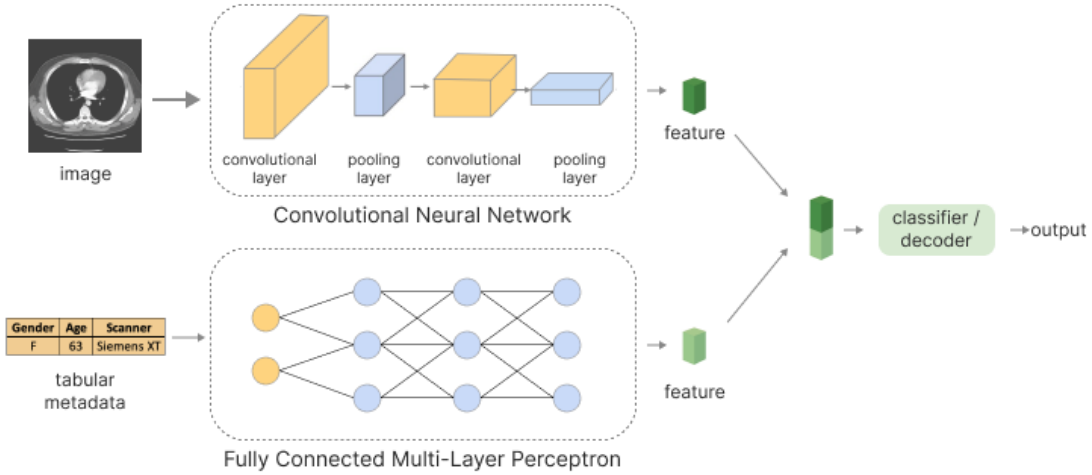


Figure 2.4: Diagram depicting example process of feature concatenation with multi-modal inputs.

Fig. 2.4 shows the image features being extracted by a CNN, whereas the tabular metadata features are extracted with a multi-layer perceptron (MLP). The different approaches are due to the different input types. CNNs are able to capture spatial information in a translationally invariant manner due to the way the convolution function works. Put simply, spatial patterns such as textures and edges can be captured no matter where they appear in the image. This also means the intermediate representations are efficient due to parameter sharing. MLPs on the other hand are very dense, as each layer is fully connected, resulting in high parametric inefficiency. This works for smaller sized inputs such as tabular data, but would be very inefficient for images - aside from this they can only accept 1D vectors as inputs.

Although this method is straightforward to implement, it does suffer significant drawbacks. The simple concatenation operation treats both modalities as the same in how it sees them, so does not effectively capture the interactions between both types. Careful preprocessing must be done to ensure the features from each modality lie in the

same ranges, otherwise there will be large weight imbalances. The ratio of input size to feature size must also be considered to ensure that one input modality is not favoured to another in its representation.

2.3.2 FiLM Conditioning Layers

Perez et al. [4] introduce an approach named Feature-Wise Linear Modulation (FiLM) as an alternative approach to incorporating metadata into a network. The key notional difference to feature concatenation is that metadata is allowed to *modulate* or *condition* network features, rather than simply being added onto them. The FiLM conditioning layers are a generalisation of conditional batch normalisation (CBN), a method which has been used in numerous applications such as speech recognition [11] and image stylisation [12].

Basic batch normalisation (BN) involves modulating mini batches of the output feature maps from layers through learned trainable scalars γ and β. For a mini-batch of feature maps $\mathcal{B} = \{\boldsymbol{\mathcal{F}}_c, ...\boldsymbol{\mathcal{F}}_n\}_{c=1}^{N}$, the batch normalised output is defined by

$$BN(\boldsymbol{\mathcal{F}}_c|\gamma_c,\beta_c) = \gamma_c \frac{\boldsymbol{\mathcal{F}}_c - E[\mathcal{B}]}{\sqrt{Var[\mathcal{B}]+\varepsilon}} + \beta_c, \tag{2.1}$$

where the parameters γ and β are learned in the optimisation process and exist to restore the representational power of the network. ε is a constant damping factor. CBN extends this method to 'ground' the features by incorporating information from metadata. γ and β are adjusted by conditioning them on a separate input vector $\mathbf{e}_m$ generated from the encoded metadata. $\Delta\gamma$ and $\Delta\beta$ are introduced as terms to modulate the learned parameters, and are generated by passing the embeddings through a multi-layer perceptron (MLP),

$$\Delta\gamma = MLP(\mathbf{e}_m) \qquad \Delta\beta = MLP(\mathbf{e}_m). \tag{2.2}$$

These predictions are then added to the learned parameters to give the final parameters used in the batch normalisation $\hat{\gamma}$ and $\hat{\beta}$,

$$\hat{\gamma}_c = \gamma_c + \Delta\gamma \qquad \hat{\beta}_c = \beta_c + \Delta\beta. \tag{2.3}$$

CBN offers a computationally efficient way of incorporating multi-modal information into a network - only two more parameters are required per layer, making the method very scalable. FiLM conditioning layers generalise the CBN method by removing the strict normalisation that precedes the affine transformation. The FiLM

generator passes the encoded metadata $\mathbf{e_m}$ through an MLP, such that it directly outputs the learned γ_c and β_c directly when given an input.

$$\gamma_c = MLP(\mathbf{e}_m) \qquad \beta_c = MLP(\mathbf{e}_m). \tag{2.4}$$

The FiLM layer then performs the feature map transformation. For convolutional networks, each feature channel is modulated by a different γ_c, β_c pair, with these values remaining consistent across spatial location within the feature map. This takes the form:

$$FiLM(\mathcal{F}_c|\gamma_c, \beta_c) = \gamma_c \mathcal{F}_c + \beta_c. \tag{2.5}$$

The authors investigate the placement of the FiLM layers, and find that their effect can be decoupled from that of the normalisation layers, with no significant improvement when placed directly after normalisation (as in CBN) compared to being placed elsewhere. Thus, the method is decoupled from the normalisation to allow for more general and versatile use. A large draw of the FiLM method is its flexibility and efficiency - with only two extra parameters per feature, a large range of interactions can be modelled between the metadata and original input. Feature concatenation however, scales with both the size of the features and number of features, without capturing *interactions* between the modalities of inputs. This dissertation will use this method due to the flexibility of being able to condition on different types of inputs - both numerical and textual in tabular format. [2]

2.4 Related Work

There has been previous research into using multi-modal inputs in the context of medical imaging. Results using the feature concatenation method of incorporating multi modal information into networks have been varied, with some reporting a drop or no change in performance, and others reporting improvements. Höhn et al use patient metadata of age, sex and anatomical site of lesion to classify skin cancer diagnoses, finding that feature concatenation is less successful than unimodally classifying the images using a CNN [13]. Here, the CNN itself renders very accurate classifications and thus the authors hypothesise that the integration of data that correlates less well with classifications would only degrade performance. Interestingly, they find that in the specific cases of low image classification confidence, replacing these decisions with a

[2] This section is partly adapted from Section 2.3 of the IPP.

unimodal metadata-only classification decision gave an overall higher accuracy. Ou et al. attempt to diagnose skin lesions with both smartphone images and 21 different clinical characteristics including age, lesion location, and lesion diameter [14]. Although they find that feature concatenation offers an improvement in model accuracy over a unimodal image classifier, their method of fusing modality features using an attention mechanism performs even better, as it is able to represent and exploit the correlations between modalities.

FiLM layers have also been used in other medical imaging applications. Lemay et al. incorporate FiLM layers into a 2D U-Net architecture to make use of information about the tumor type during spinal cord tumor segmentation, and also make use of organ type in multi-organ segmentation [15]. They find that the use of this metadata in the spinal cord segmentation task offers a 5.1% improvement in Dice scores, and that the multi-class FiLMed network performs comparably to the single-class U-Net. Chartsias et al. use FiLM layers in only the decoder half of their architecture, in order to separate out disentangled representations of cardiac images, by modality factors and anatomical factors [16]. They show these representations to be useful in synthesising CT from MRI, and vice versa. Jacenków et al. incorporate FiLM layers preceded by a feature-wise attention mechanism to incorporate spatial information into a cardiac structure segmentation task, naming this method INSIDE [17]. The spatial information consists of slice location, and cardiac phase. They found that although the INSIDE method always performed as good as, or slightly better than the baseline, it offered the largest improvement when training with smaller subsets of the data.[3]

As with the feature concatenation method, findings are inconsistent, with some authors concluding that FiLM layers do not always offer improvements. Vincent et al. incorporate MRI contrast as metadata using FiLM layers, and find that this offers no improvement in Dice score compared to a well optimised U-Net implementation [18]. Sheth et al. investigate the use of CBN, a variation of FiLM layers combined with normalisation (see Section 2.2.2), and find that in certain multi-modal tasks the visual features alone without metadata are superior for a generalisable model [19]. They evaluate this through the task of tumour type classification of histology images, with additional metadata including age, gender, and size of tumor cells. They find that although CBN outperforms simple BN when using this metadata as conditioning input, it actually encourages the network to learn less representative visual features. The classification accuracy when using *only* metadata as input to classify tumour type is

[3]This paragraph is taken from Section 3 of the IPP.

already so high, that using it to condition an image-based CNN encourages the network to learn shortcuts between metadata and labels. Although overall accuracy is high, the resulting network is one that generalises poorly.

A key gap in the research that this dissertation will attempt to address is a lack of clear comparisons between incorporating different *types* of metadata. Most authors either include all available metadata [13, 14, 20], or only use one source [15, 18]. This dissertation will go further in comparing the effect of incorporating different types of metadata by using the FiLM method. By doing so, the link between the relevance of metadata, and effectiveness of incorporating it can be established more thoroughly.

Chapter 3

Methodology

3.1 ATLAS Dataset

The dataset explored in this research will be the publicly available ATLAS dataset R2.0 [3]. This version of the dataset contains 655 manually annotated T1-weighted MRI scans with corresponding metadata, with no normal (healthy) scans present. T1-weighted MRI enhances the signal of fatty tissue, whilst suppressing that of water [21], and is the most effective MRI modality in showing post-acute infarcts. The dataset is collected from 44 different research cohorts, with every scan corresponding to the first timepointed scan since stroke, and thus each scan being of a unique patient. Each scan is annotated by an expert radiologist, with a second radiologist then performing a quality control check across all manual segmentations.

3.1.1 Existing Dataset Metadata

The metadata provided with the dataset contains the following:

- **Lesion Characteristics** - This includes the **laterality** and **location** (anatomical region) of the lesion. Laterality, referred to in the dataset metadata as 'Stroke Hemisphere', defines the side of the primary stroke location, and can belong to 'Left', 'Right' or 'Other' ('Other' pertaining to central regions such as in the brainstem and cerebellum). Location provides the precise neurological region of the lesion. It includes regions such as the occipital lobe, temporal lobe and cerebellum (see *Table 3.1*). Both are given 'Primary' and 'Secondary' entries to denote the characteristics of primary and secondary lesions if more than one is present.

- **Chronicity** - This is the number of days post stroke that the MRI is acquired. For some scans, the exact number is not available and the only information is that the scan occurred over 180 days post stroke.
- **Scanner** - This is the scanner model type.

The labels available for Primary/Secondary Stroke Hemisphere, Primary/Secondary Stroke Location and Scanner are shown in *Table 3.1*. It should be noted that for Primary/Secondary Stroke Hemisphere and Scanner, labels are mutually exclusive whereas for Stroke Location multiple labels can be assigned.

Scanner	Primary/Secondary Stroke Hemisphere	Primary/Secondary Stroke Location
GE 750 Discovery	Right	Basal Ganglia
GE Signa Excite	Left	Brainstem
GE Signa HD-X	Other	Brainstem/Pons
Philips		Caudate
Philips Achieva		Cerebellum
Siemens Allegra		Frontal Lobe
Siemens Magnetom Skyra		Hippocampus
Siemens Prisma		Insula
Siemens Skyra		Occipital Lobe
Siemens Sonata		Pallidum
Siemens Trio		Parietal Lobe
Siemens TrioTim		Putamen
Siemens Verio		Temporal Lobe
Siemens Vision		Thalamus

Table 3.1: Table showing all possible values categorical metadata types can take.

For this research to be well motivated, it must be clinically feasible that this metadata can be gained from just the patient, or the scan, so that it can then be used in an automatic segmentation model. In the case of laterality, this can be inferred from the patient symptoms. In general, strokes affecting one side of the brain can lead to symptoms on the opposite side of the body as the brain's hemispheres control opposite sides of the body. Thus, significant paralysis or weakness in the right side of the body with moderate weakness in the left side would imply a primary **left** stroke hemisphere and secondary **right** stroke hemisphere. Similarly, stroke location can also be inferred from patient symptoms.

Fig. 3.1 shows different regions of the brain with the human function they correspond to. It then follows that a degradation in specific functions can map to specific

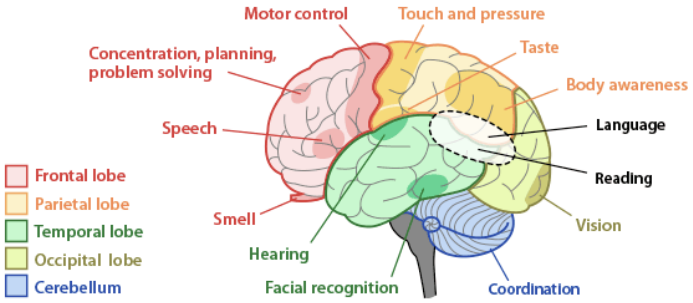


Figure 3.1: Diagram showing anatomical regions of the brain with the associated human trait they correspond to.

lesion locations. However, as well as the link between lesion characteristics and symptoms, ultimately it is true that even if a radiologist is required to provide this high level metadata, there would still be a considerable time saving compared to manually segmenting a lesion slice by slice, and thus the motivation for incorporating this information into the automatic segmentation task holds.

3.1.2 Derived Metadata

Due to the mixed results and previous lack of success in using FiLM layers for segmentation [17, 22], as well as using the metadata provided in the ATLAS dataset, this thesis will investigate the use of synthesised metadata which is derived directly from ground truth. It is hoped that incorporating this information will help to establish an upper limit of how 'useful' metadata must be in relation to the images for the FiLM method to prove successful. All derived values have a clinical link and justification - either they are correlated to information that is readily available when imaging stroke, or they are linked to information that can be easily marked by a clinician.

Derived values include:

- **Mean Lesion Pixel Intensity -** This is calculated by masking the normalised MRI image with the ground truth array, and then calculating the mean pixel intensity of the masked array. The link here is with the readily available 'Chronicity' information. Stroke images acquired in later chronic phases have a different appearance on T1-weighted images. As the tissue in the affected area undergoes atrophy and degeneration, the signal intensity on T1-weighted images decreases, causing the lesion to appear hypointense (darker). Newer ischaemic lesions which occur in the acute phase of a stroke (within first 48 hours), typically appear as

hyperintense (bright) on T1-weighted images. This is because the tissue in the affected area is still swollen and contains excess water, which increases the signal intensity on T1-weighted images. This difference in appearance can be seen in *Fig. 3.2*.

- **Lesion Centre of Mass -** This is calculated by finding the centre of mass of all labelled lesion pixels as (x, y, z) coordinates. It is plausible that a centre of mass value could easily be generated from a radiologist drawing rough bounding boxes around lesions, rather than marking them pixel by pixel. A downside of using this value is that it does not take into account the different centres of mass of separate lesions.

- **Lesion Volume -** This is provided in the dataset, but treated as a derived value as it is derived directly from the ground truth. It is given as an integer in terms of voxels cubed. Exact lesion volume is not possible to be inferred without an exact segmentation, however, if it can be shown that incorporating the numerical values have a significantly positive effect, further experiments can be run with stratified labels (Very Small, Small, Medium etc.), as it is plausible that a radiologist could mark these reasonably quickly.

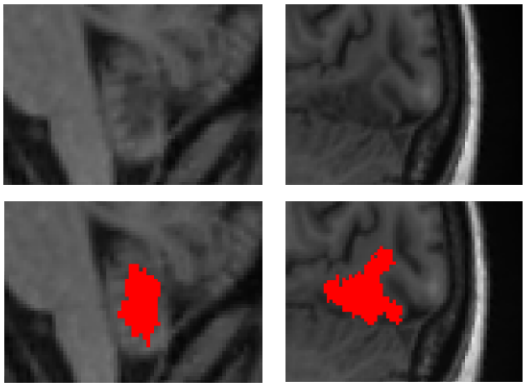


Figure 3.2: Acute lesion (left) on MRI taken 1 day after stroke onset and chronic lesion (right) on MRI taken 670 days after stroke onset. Note lower intensity for chronic lesion.

Table 3.2 shows a summary of all investigated MRI metadata, with their data type, ranges if continuous, and number of classes if categorical. Class labels can be found in *Table 3.1* Derived metadata values are shown by the shaded rows. Although Lesion Volume is provided with the dataset, it is treated from herein as a derived value, as it can only be derived directly from a segmentation and not independently of it.

Metadata	Type	Range	Nr. of Categories
Primary Stroke Hemisphere	Categorical	-	3
Secondary Stroke Hemisphere	Categorical	-	3
Primary Stroke Location	Multi-Label Categorical	-	14
Secondary Stroke Location	Multi-Label Categorical	-	12
Scanner Type	Categorical	-	14
Lesion Volume	Continuous	13 - 496656 voxels	-
Chronicity	Continuous	1 - 10806 days	-
Mean Lesion Pixel Intensity	Continuous	-0.635 - 2.040	-
Lesion Centre of Mass	Continuous List	(14, 21,7) - (71,85,71)	-

Table 3.2: Table showing all types of investigated metadata, synthesised metadata and their corresponding ranges/categories. Lesion centre of mass is given for mris resampled to an image size of 88x112x96 pixels, as used to train all models. Derived values are shown by the shaded rows.

3.1.3 Preliminary Metadata Analysis

The lesion laterality metadata is first investigated to verify how accurate and useful the labels inherently are. To do so, the ground truth arrays are summed through the coronal axis for all volumes in the training and validation sets for each pair of Primary Stroke Hemisphere and Secondary Stroke Hemisphere values. Dataset images are already registered to the MNI-152 template, so no further registration is required to verify laterality. Results are shown in *Fig. 3.3*.

We can see lesions tend to be concentrated in the frontal half of the brain when given ‘Right’ or ‘Left’ labels for laterality, and more in the occipital area when given the ‘Other’ label. Interestingly, for lesions labelled ‘Other’, there seems to be a bias towards left-sided lesions. *Fig. 3.3* verifies that the ground truths are well separated for different lateralities, with the most variation between volumes occurring when the Primary and Secondary lateralities do not agree. A Primary stroke hemisphere label of ‘Other’ is potentially the most informative, as the ground truths across all volumes appear concentrated within the smallest area - around the Brainstem or Cerebellum, thus it is hypothesised this label would provide the most useful information to the network.

The distribution of the two continuous metadata variables, chronicity and lesion volume, is shown in *Fig. 3.4*. They follow a similar distribution, with values concentrated on the lower end, and the upper quartile of values having a very large range. If there are not enough examples of large or late imaged lesions in the training set, it could be that

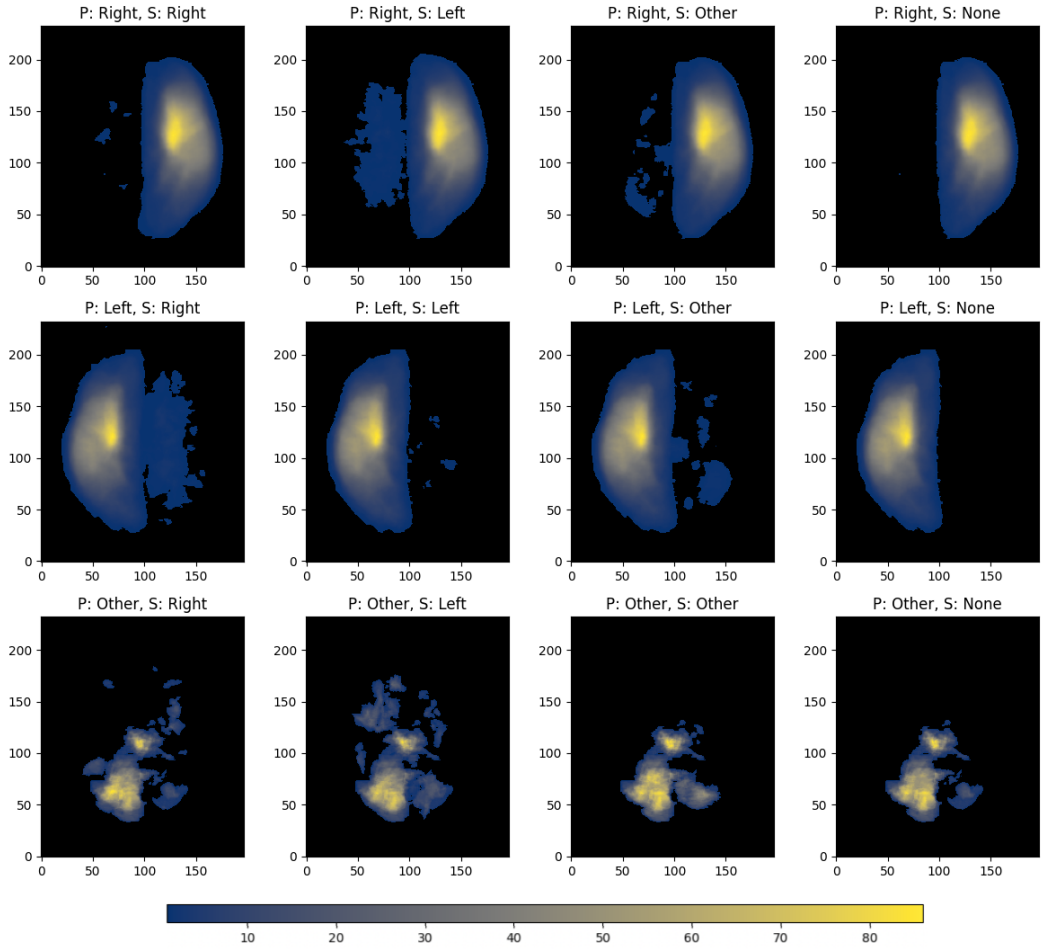
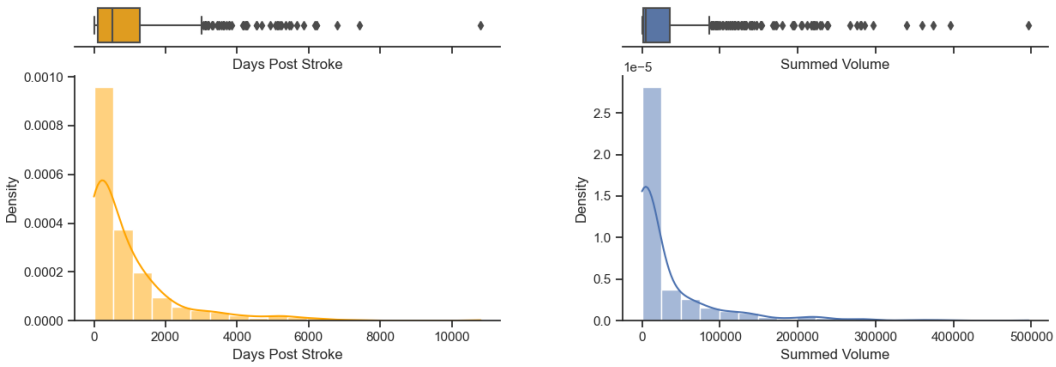


Figure 3.3: Colour denotes number of pixels containing lesion at that location, summed along the dorsal/ventral axis for subsets of volumes with different Primary (P) and Secondary (S) Stroke Hemisphere label combinations. All plots are in RAS+ space at a resolution of 2x2x2 mm/pixel.

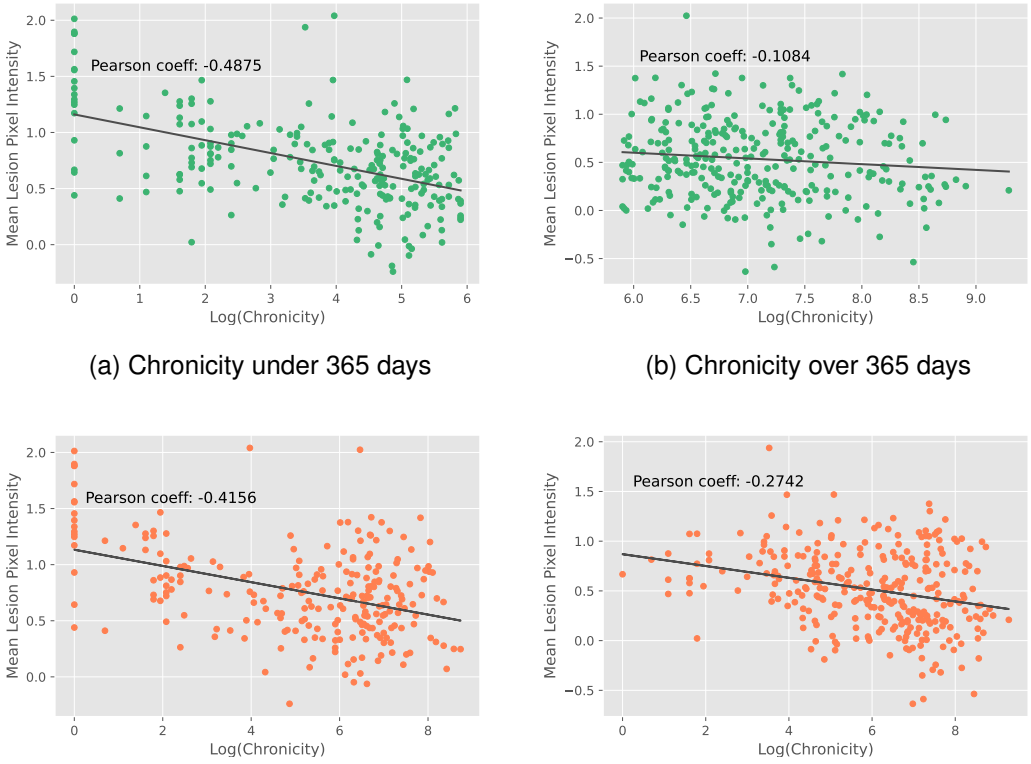
this information proves arbitrary to condition the network with.

The correlation between chronicity and mean pixel intensity is investigated by plotting them against each other for different subsets of chronicity and lesion volume values- *Fig. 3.5*. The hyperintense acute lesions (within 24 hours of stroke onset) can clearly be seen in *Fig. 3.5a*, clustering in a vertical line at the origin. The trend is much stronger for both images with smaller lesions, and those which have been acquired within a year of stroke onset, shown by the higher Pearson coefficients for these subsets. This should be considered when analysing the results of incorporating Mean Lesion Pixel Intensity using conditioning layers.



(a) Chronicity (b) Lesion Volume

Figure 3.4: Boxplot and histogram of Chronicity and Lesion Volume values for every volume in training set. Mean value of Chronicity is 511 days, mean value of Lesion Volume is 4264 voxels cubed.



(a) Chronicity under 365 days (b) Chronicity over 365 days

(c) Lesion Volume below 4264 voxels (median) (d) Lesion Volume above 4264 voxels (median)

Figure 3.5: Trends between Chronicity and Mean Pixel Intensity for different dataset splits of Chronicity and Lesion Volume values. Mean Lesion Pixel Intensity is plotted as a standardised value.

3.1.4 Metadata Encoding and Preprocessing

Metadata is provided with the dataset in tabular format, with a mix of categorical text values and numerical continuous values. Neural networks require a numerical input, therefore the categorical values must first be numerically encoded and then concatenated with continuous values into a single vector. One-hot encoding is used for categorical data to ensure that categorical variables are represented in a meaningful way. This converts each categorical value into a binary vector of 1s and 0s with a separate binary vector created for each unique category. This representation prevents the algorithm from assigning any inherent order or magnitude to the categories. A downside of one-hot encoding can be that dimensionality of data greatly increases for data with many distinct categories, however as seen in *Table 3.2*, the maximum number of categories for any one metadata variable is 14 which is of negligible size compared to the size of the image inputs (88x112x96).

NA values when using one-hot encoding are easily dealt with, with 0 values in each unique category binary vector. The only categorical metadata types which have NA values are Secondary Stroke Hemisphere and Secondary Stroke location when only a single lesion is present in the ground truth. Thus the encoding of all zeros signals this to the network. For the continuous metadata types, NA values must be manually filled. Centre of Mass and Mean Lesion Intensity are calculated directly from ground truth, and thus have no missing values. Lesion Volume NA values are filled by deriving these directly from the ground truth. Lastly, Chronicity NA values are filled in using the mean value, which is 511 days. Continuous values are standardised to a mean of 0 and standard deviation of 1. Standardising the features ensures that each feature contributes equally to the learning process, otherwise features with larger scales would dominate the learning process. Standardization helps this by creating a more balanced optimization landscape, allowing the algorithm to converge more quickly and reliably. It is chosen over normalisation (where values are normalised to lie in the range [0,1]) due to the distributions observed in *Fig. 3.4*. If normalisation was chosen, many values concentrated on the lower end of the scales would be normalised to zero, just to accommodate the few outliers at the upper end of the scale.

3.2 Model Architecture

3.2.1 FiLMed U-Net Model

The baseline model used for experiments is a standard U-Net [1] (See Chapter 2.2). Metadata is then encoded as in Chapter 3.1.4, and input to the network through the FiLM generator which generates the parameters β_c and γ_c to modulate the feature maps by. The full architecture is shown in *Fig. 3.6* for a 'late fusion' version of the architecture wherein FiLM layers are only placed in the decoder half of the network. The decoder is responsible for generating the final segmentation, whereas the encoder is responsible for generating latent representations of the image. By applying FiLM conditioning in the decoder only, the model can adapt its behavior based on the given context, without affecting the lower-level features learned by the encoder. This is even more justified for the U-Net as opposed to a standard autoencoder due to the skip connections - unmodulated features from the encoder are concatenated with decoder features before passing through FiLM layers rather than already modulated features being concatenated and then remodulated through another FiLM layer. Both late fusion and complete fusion, where FiLM layers are placed at every network layer, will be investigated to see if certain types of metadata work better with a certain method.

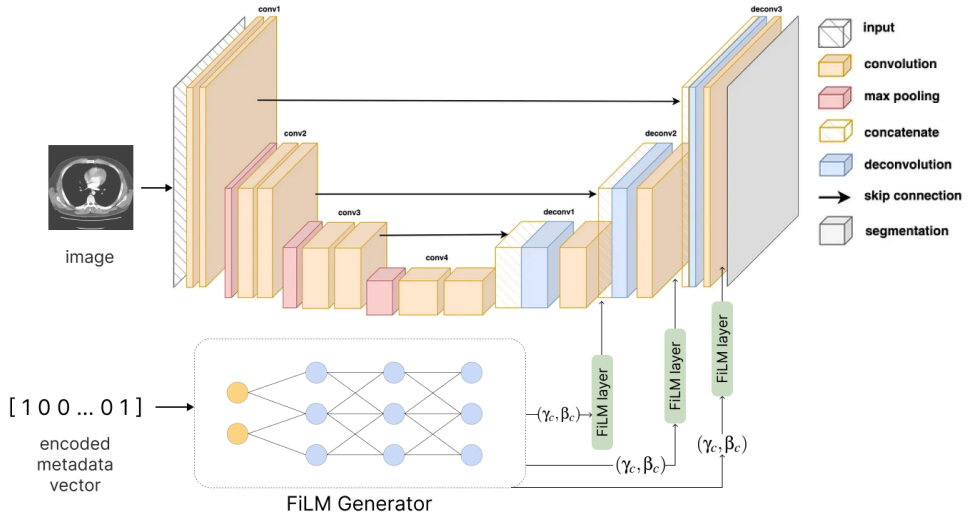


Figure 3.6: Diagram showing a FiLMed U-Net. To demonstrate how metadata can be injected at different locations, late fusion is depicted where FiLM layers are only placed in the decoder half of the network. Base diagram from [23].

3.2.1.1 Training Details

Training is conducted on a machine equipped with an Intel Core i9 processor (3.5 GHz, 12 cores), 64 GB of RAM, and an NVIDIA GeForce RTX 2080 Ti GPU. The scripts used to create the model and run training are written in Python and PyTorch version 2.0.1. Model hyperparameters used for model training are shown in *Table. 3.3*, and are optimised to achieve the best baseline performance. Batch size is maximised for training with the dataset on the above machine. A grid search is conducted for learning rates between $1e-5$ and $1e-1$ to find the best value, and then a one cycle scheduler is added to modulate the learning rate over the course of training. Models are initially run for 150 epochs to find that around the 60th epoch, validation scores begin to plateau, therefore 75 epochs is chosen as the total training time. For every new model that is run, training curves are visually assessed to ensure that validation scores also plateau around the same time, in case metadata models take longer to converge.

The loss function chosen to optimise the network with is an average of Dice Loss and Binary Cross Entropy (BCE) Loss. Dice Loss is calculated as 1 - Dice Score (see Chapter 3.3.1). The formula for calculating the BCE loss for a single pixel is:

$$L_{BCE} = -\left(y\log(p) + (1-y)\log(1-p)\right) \tag{3.1}$$

where y is a binary indicator indicating whether the predicted class is correct or not, and p is the probability of the pixel lying in the correct class. The loss is calculated for every pixel in an image, and then an average taken.

As BCE loss is evaluated on individual pixels, the misclassification of larger lesions will contribute more than small ones. Dice loss suffers the opposite problem as it is calculated on a per-volume basis; a small misclassification in a small lesion will result in the same loss as a very large misclassification in a larger lesion. Hence, the losses are averaged to counteract the limitations of one another.

3.2.2 Metadata Prediction Model

In order to investigate how much information from each type of metadata is already encoded in the images alone, the task is reversed to see if there is a link between the ability to infer the metadata already from the images using a CNN, and the performance increase by injecting this same metadata using FiLM. The aim is to establish how ‘new’ each type of metadata is to the model, and thus how useful different types *should* be to condition the network with. It is hypothesised that if the model is able to predict a

Hyperparameter	Value
Network Depth	4
Batch Size	3
Learning Rate Scheduler	OneCycleLR
Max Learning Rate	0.01
Optimiser	AdamW
Training Epochs	75
Image Size	(88, 112, 96)
Loss Function	Dice and BCE Loss

Table 3.3: Training hyperparameters for metadata FiLMed models.

metadata type with incredibly high accuracy already, it will not be useful to condition the network with, as the information is already encoded in the images. To do so, a model architecture is used that utilises the encoder half of the U-Net with a linear layer as the final layer, trained such that the compressed representation is now the encoded metadata vector - shown in *Fig. 3.7*. The prediction of each individual metadata type is treated as a separate task, with different models trained to predict each one.

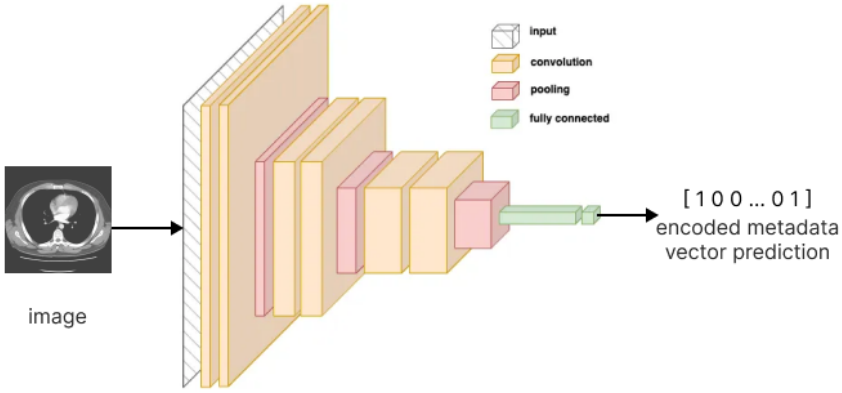


Figure 3.7: Diagram showing architecture used for metadata prediction. Base diagram from [23].

For predicting continuous metadata values, Mean Squared Error (MSE) is used as a loss function, with R^2 Score (regression coefficient) as an evaluation metric. The R^2 Score quantifies the proportion of variance in the target variable (in this case the ground truth metadata), that is explained by the inputs (in this case the model metadata predictions). A score of 1 would indicate the predictions perfectly match ground truth, and 0 implies no correlation.

For predicting categorical metadata values, Binary Cross Entropy Loss (*Eqn. 3.1*) is used to optimise the model, and F1 Score is used to evaluate predictions. Accuracy is a non-optimal metric to use for this, as in class-imbalances cases, high accuracy can be achieved by simply predicting the majority class for all instances whilst performing poorly on the minority class. As the metadata values are one-hot encoded, all input values are binary, and thus both micro and macro averaging for F1 Score give the same results. The same hyperparameters are used as in *Table 3.3*, except for the aforementioned loss functions and the learning rate. When conducting a learning rate grid search, a value of $1e-3$ is found to be optimal.

3.3 Experiments and Evaluation

Experiments will be conducted both by varying the type of metadata the network is conditioned with, and whether complete fusion (FiLM at every layer) or late fusion (FiLM in decoder only) is used. This way, the effect of each individual type of metadata can be evaluated and compared. A model incorporating all metadata, and also random noise as metadata, will also be trained for further comparison. The baseline is taken to be a simple U-Net model with no FiLM conditioning or metadata input.

Each model is evaluated by performing 3 cross validation runs with separate validation sets, split 80:20 training to validation. This gives 524 volumes in each training set, and 131 in each validation set. For every run, the Dice score (see Chapter 3.3.1) for the best performing model over all epochs on the validation set is taken. These 3 values are then averaged, with the standard deviation also calculated to be able to compare the significance of differences between model runs.

3.3.1 Dice Score

To evaluate the accuracy of model predictions compared to ground truth for each volume in the validation set, the Sørensen–Dice coefficient (herein referred to as the Dice score) is used. The formula is given as such:

$$D = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{3.2}$$

where Y is the ground truth array and $\hat{Y}$ is the binarised (i.e. thresholded) model segmentation prediction. Thus, $|Y \cap \hat{Y}|$ represents the intersection of the two arrays, and $|Y| + |\hat{Y}|$ the sum of the volumes of both ground truth and prediction - this is not

the same as the union of both arrays. The Dice score is the most common metric used in image segmentation, as it essentially measures the similarity between two sets. A limitation of using the Dice coefficient to evaluation segmentation performance is its variance in evaluating errors for small and large segmentations. A single pixel error in a small segmentation has the same effect as completely omitting a large lesion, which may not be desired behaviour. The Dice score is calculated separately for every volume in the validation set and averaged to get a representative value for model performance.

3.3.2 Decision Trees

The Dice score is useful as a high level metric to compare the average model performance across all lesions, however we would also like to investigate the characteristics of the individual cases where metadata FiLMed models exhibit superior predictive performance. Some authors have used decision tree analysis for CNN prediction analysis in order to increase the explainability of the learned visual features of the CNN [24, 25]. However, since we have access to a set of features pertaining to the dataset already - the metadata - we can use decision tree regression to predict the performance of a model across a dataset (using Dice scores), given only the metadata. This will expose specific metadata characteristics that result in FiLMed model improvements or deterioration compared to the baseline. Of course, this could be inferred by manually inspecting all predictions, however decision tree analysis gives us a more efficient and visual way to do this.

Decision tree regression is a machine learning algorithm that can be used for predicting continuous values based on a set of input features. For this specific analysis, we can take the input features to be the encoded metadata vector for each image, and the continuous variable to predict as the Dice score from a particular model's segmentation prediction, or the *improvement* in Dice score from baseline. The algorithm constructs a tree-like structure, where each node represents a decision based on a specific feature's value, and each leaf node represents a predicted numeric value for the target variable. The goal is to divide the feature space into regions that correspond to different target values. This visualisation will give clear insight into the metadata characteristics that result in certain predictions having higher or lower Dice scores.

Decision tree regression works by recursively dividing the dataset based on the value of one feature at a time, whilst enforcing similarity between the target values in each subset of the divide. The most common splitting criteria is mean squared error,

which measures the average squared difference between predicted and actual values. To run inference and make a prediction, the algorithm traverses the tree depending on the feature values and decisions at each node until reaching a leaf node, where the value is used as the regression output.

For implementation, the open source scikit-learn `DecisionTreeRegression` class is used. The max tree depth is set to 3 for interpretability, and minimum number of samples in each leaf node to 3 to combat against overfitting. Mean squared error is used to optimise the node splits. *Fig. 3.8* shows an example of a regression decision tree.

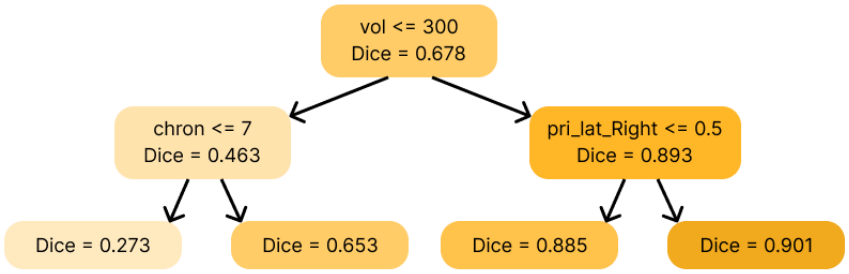


Figure 3.8: Example decision tree. Top line in each node shows the feature and feature value the node is split on. Value of the node shows the average Dice score for all values falling within the subset of samples beneath it. Saturation of node colour corresponds to magnitude of predicted Dice score. In this example, Right Primary Stroke Hemisphere lesions over a volume of 300 would receive the most accurate segmentations from the model being evaluated.

Chapter 4

Results and Analysis

4.1 Metadata FiLMed Model Results

Table 4.1 shows the average Dice scores for all trained metadata FiLMed models. Best models are indicated in bold for late fusion, complete fusion, and also for single metadata and multiple metadata models. As well as the baseline model with no metadata or FiLM layers, a model is also trained with a length 30 random binary vector attached to each input MRI as metadata. This acts as another type of baseline, to verify that the metadata FiLMed models are learning representative information from the conditioning rather than any increase in performance being due to added model complexity.

The best performing model is one that incorporates all available metadata (including derived values), using complete fusion - where FiLM layers are placed at every layer of the network, and offers on average a 7.3% absolute increase in performance over the baseline model. Incorporating stroke location with complete fusion gives the best result for non-derived metadata, with a 6.3% absolute increase in performance. Interestingly, all models incorporating only one type of metadata, except the Lateralities model, showed better performance with a late fusion architecture. Models incorporating multiple types of metadata however, show better performance when using complete fusion. Injecting multiple types of metadata might require the network to have the flexibility to modulate *all* feature maps throughout the network, rather than only those in the decoder, so as to be able to represent the complex relationships between metadata values themselves as well as between the metadata and MRIs - *Fig. 3.5* demonstrates the non-linear relationship between different metadata values.

The best three non-derived metadata models are Locations with Complete Fusion, Lateralities and Locations with Complete Fusion, and Chronicity with Late Fusion.

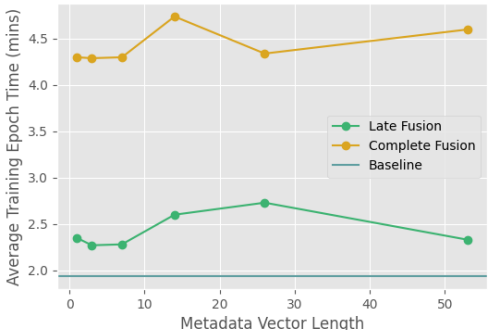
	No Fusion	Late Fusion	Complete Fusion
No Metadata	0.5039 ± 0.0047		
Binary Noise		0.5059 ± 0.012	0.5061 ± 0.013
Centre of Mass		0.5183 ± 0.0071	0.5112 ± 0.0048
Lateralities		0.5240 ± 0.0055	0.5136 ± 0.0051
Mean Intensity		**0.5315 ± 0.0029**	0.5243 ± 0.0148
Volume		0.5229 ± 0.0026	0.5087 ± 0.0088
Locations		0.5257 ± 0.0089	**0.5356 ± 0.0157**
Chronicity		0.5294 ± 0.0089	0.5229 ± 0.0109
Scanner		0.5154 ± 0.0019	0.4832 ± 0.036
Lats and Locs		0.5280 ± 0.0059	0.5352 ± 0.0076
All Non Derived Met		0.5159 ± 0.0018	0.5248 ± 0.0105
All Metadata		**0.5316 ± 0.0066**	**0.5409 ± 0.0108**

Table 4.1: Table of results for all experiments run with inserting different types of metadata to condition the U-Net. Value shows the average dice score across 3 cross validation splits, with the standard deviation between runs also shown. Late fusion pertains to models where FiLM layers are inserted at every level of the decoder half of the network, whereas in complete fusion FiLM layers are inserted at every level in both encoder and decoder halves. Models using derived metadata have been highlighted in yellow. Models using more than one metadata type are shown below the bold line. Top left entry shows the baseline result with no metadata and no FiLM layers.

FiLM layers do not directly encode spatial information, as each individual feature map is given the same affine transformation across the map. However, these results show that this spatial metadata can still provide new information to the network, implying that lesions present visually differently, and distinctly, in different areas of the brain.

There also appears to be a trend where complete fusion models display slightly more variance between runs than using late fusion. This can be seen in *Table 4.1* most prominently for the Mean Intensity, All Metadata and Location FiLMed models. This is potentially due to a combination of the increased model complexity and also small validation set size. It is possible that the extra parameters from the conditioning layers are able to capture slightly more specific patterns in the training data and overfitting to these, leading to better performance when these patterns are also present in the validation set, and worse when they are not - resulting in higher variance between data splits.

Model complexity should also be considered with regards to model run times, shown in *Fig. 4.1*. Average epoch time is plotted against different input encoded metadata vector lengths for late and continuous fusion models, as well as the average total training time shown for different model types. This illustrates that the model complexity, and training time, scale with the number of FiLM layers rather than the size of the input metadata. Although the two best performing models use complete fusion, when taking into account the variance in performance between model runs, and total training times, it is perhaps not always the superior method. Late fusion offers a better tradeoff between increase in performance, and run time.

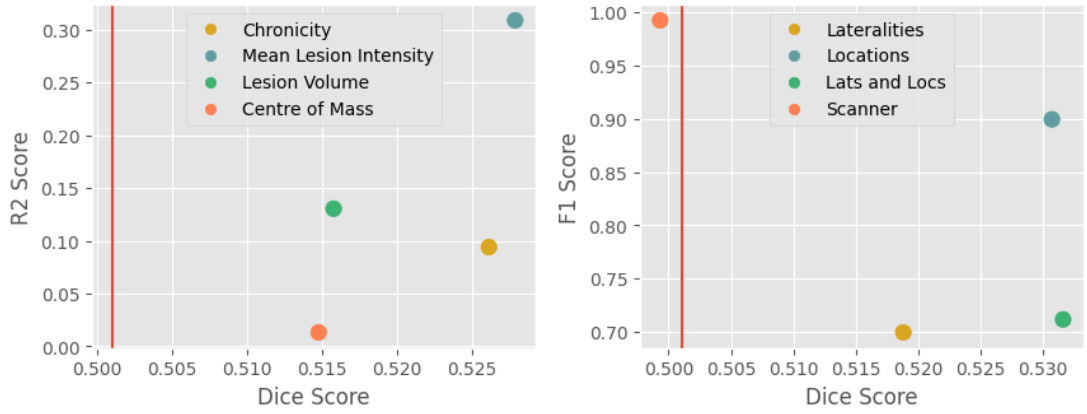


Model Type	Avg. Training Time (mins)
Baseline	145.35
Late Fusion	182.00
Complete Fusion	332.13

Figure 4.1: Left plot shows average training epoch time for models trained with different lengths of metadata conditioning vectors. Right table shows average training time over 75 epochs for a single run, for all three model architectures.

4.2 Metadata Prediction

Results for predicting metadata from images alone are shown in *Fig. 4.2*, with separate graphs denoting prediction of continuous metadata and prediction of categorical metadata. Prediction metrics are plotted against the average Dice score achieved when conditioning models with that same metadata, to establish whether there is a link between the effect of incorporating metadata for automatic lesion segmentation, and how much of that metadata is already encoded in the MRI images themselves. It should be noted that the correlations lie on very different scales - in *Fig. 4.2b*, R^2 scores for continuous metadata are all below 0.35. This shows low predictive performance and implies that none of these values are encoded in the MRIs already, and thus should improve the network when injected as metadata as they are 'new' and distinct information - which we can see they do. In *Fig. 4.2b*, F1 scores are all above 0.7, showing high correlation between the categorical metadata and images themselves. In the most extreme case of the Scanner type which can be predicted almost perfectly, this in fact has the opposite effect when injected into a FiLMed model, offering a decrease in performance over baseline. However all other categorical metadata values which are predicted with high accuracy, but not perfectly, still offer a large improvement over the baseline Dice when used as conditioning inputs to FiLMed models.



(a) Continuous Metadata

(b) Categorical Metadata

Figure 4.2: The average Dice score achieved when incorporating different metadata types is plotted against prediction performance when predicting the corresponding metadata value from the MRI only. Baseline dice is shown as a vertical line for comparison.

4.3 Analysis of Trained Models

4.3.1 Laterality Models

This section draws on the preliminary analysis performed in Chapter 3.1.3 to investigate the effect of incorporating stroke laterality into a FiLMed model on the resulting predictions. Predictions on the validation set from the best performing Laterality FiLMed model are chosen to analyse. Firstly, a kernel density estimation plot in *Fig. 4.3* shows the distribution of Dice scores across the validation set for both the baseline and the Laterality FiLMed model. Here we can clearly see a shifting of mass away from near zero scores for the Laterality FiLMed model, showing that the incorporation of Laterality metadata specifically helps in cases where lesions are almost entirely missed. For volumes where Dice score is already very high however, incorporating this additional metadata has no effect, which is to be expected. If the Dice score is already greater than 0.8, it is likely that the laterality of the lesion has already been predicted correctly and this conditioning metadata would offer no extra information.

These Dice scores are then taken and stratified by their 'Primary Stroke Hemisphere' label to find the average Dice score for predictions for each label. Results are shown in *Table 4.2*. A notable increase in performance is shown for volumes with a primary stroke hemisphere labelled 'Other'. It should be noted that this subset is smaller than subsets with labels 'Right' and 'Left' (16, 60 and 54 respectively), however the subset contains enough volumes for the result to be significant.

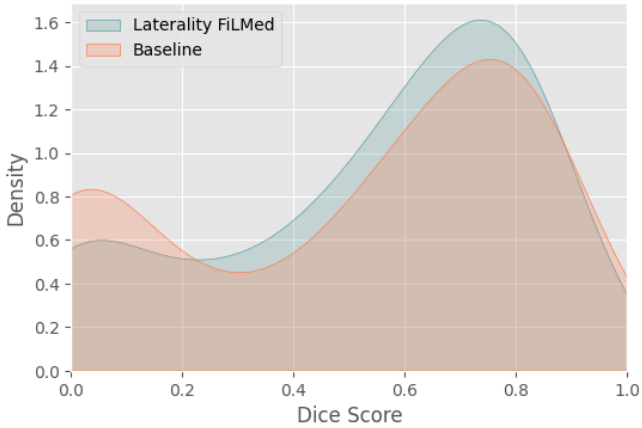


Figure 4.3: Kernel density estimation plot for Dice scores over validation set for both baseline and Laterality FiLMed models.

To inspect this further, predictions for the same laterality subsets for both models are

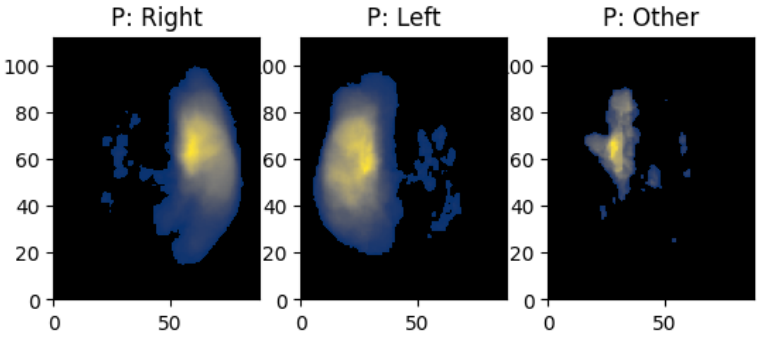
Model	Primary Stroke Hemisphere		
	Right	Left	Other
Baseline	0.5526	0.5385	0.0947
Laterality FiLMed	0.5693	0.5402	**0.2812**

Table 4.2: Dice scores across one validation set, split by 'Primary Stroke Hemisphere' labels. Results for baseline model, and Laterality FiLMed late fusion model are shown. There is a notable increase in Dice scores for volumes labelled 'Other' when incorporating laterality as conditioning metadata.

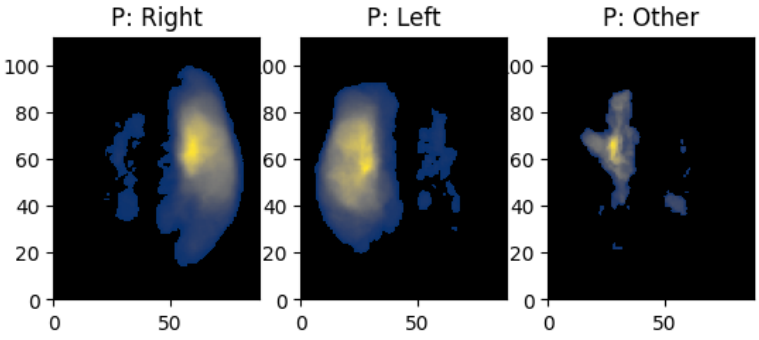
visualised in *Fig. 4.4*. As in *Fig. 3.3* in the preliminary analysis, predictions are summed through the dorsal/ventral axis to verify at a high level if the lesions are being predicted with the correct laterality. Even though *Table 4.2* reports an increase in performance for models with laterality label 'Other', it is hard to visually see this between the laterality FiLMed model and baseline predictions in *Fig. 4.4*. Both models seem to completely miss lesions in the Cerebellum region are of the brain, hence the overall low Dice scores for 'Other' labelled laterality volumes compared to 'Right' and 'Left'. Ultimately, the translational invariance of FiLM conditioning means that spatial metadata can only have an effect if this spatial information also encodes visual features, which in the laterality case may not be true.

4.3.2 Decision Tree Analysis

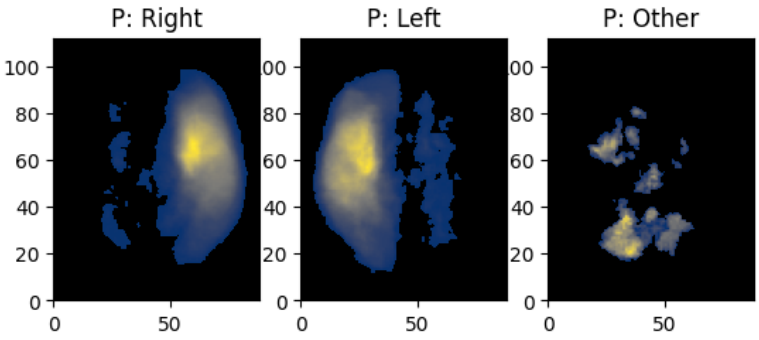
Decision trees are used to analyse the predictive performance of different models across the metadata features of a validation set. The baseline model is first considered, by using an encoded vector of all metadata values to predict the baseline dice. The result is shown in *Fig. 4.5a*. At every level of the tree, volume is shown to be one of the splitting features, showing it is a strong predictor of Dice score. This is more of a reflection of the bias in using Dice score as a metric, than it is the feature importance of lesion volume. As discussed in Chapter 3.3.1, the Dice metric penalises small errors in both small and large volumes differently. As larger lesions have a lower surface area to volume ratio, boundary errors result in a lower decrease in Dice score than boundary errors for a small lesion. Hence the observed behaviour in *Fig. 4.5a* - where the highest average Dice scores are seen for the largest lesions. Therefore in subsequent analysis, lesion volume is removed from the input metadata vector to give a more representative



(a) Laterality FiLMed Model Predictions



(b) Baseline Predictions

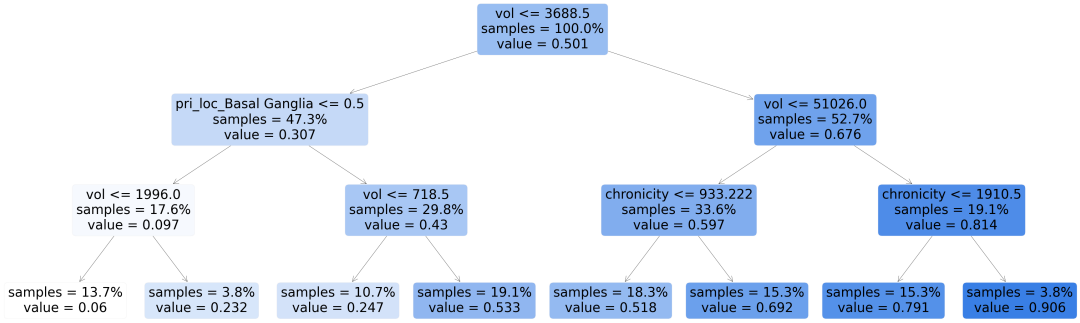


(c) Ground Truth

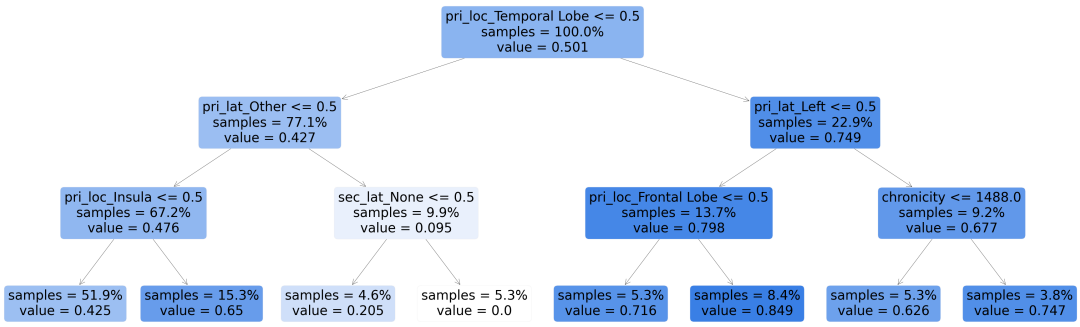
Figure 4.4: Segmentations summed through dorsal/ventral axis for subsets of 'Primary Stroke Hemisphere' (P) labels. Top row shows predictions from a FiLMed model using both Lateralities as metadata conditioning input, middle row shows the baseline predictions, and the bottom row shows the ground truth.

and less biased metadata feature analysis.

Fig. 4.5b shows the same model architecture but now trained with volume removed from the input metadata vector. Here we can see confirmed the behaviour observed in Chapter 4.3.1, with poor performance for volumes with primary laterality 'Other' (observed in the right side subset below the left node marked 'pri_lat_Other ≤ 0.5').



(a) Baseline Model

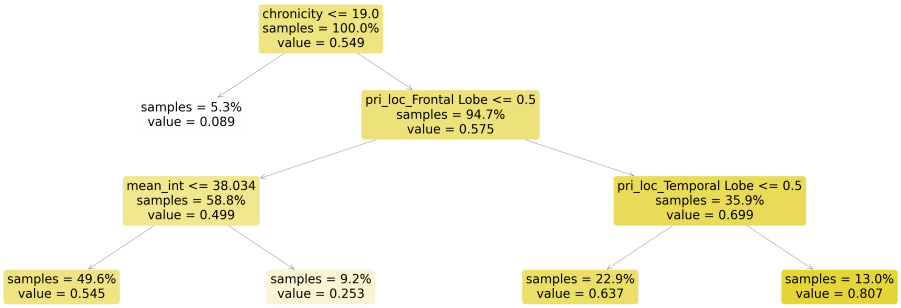


(b) Baseline Model, Vol Removed

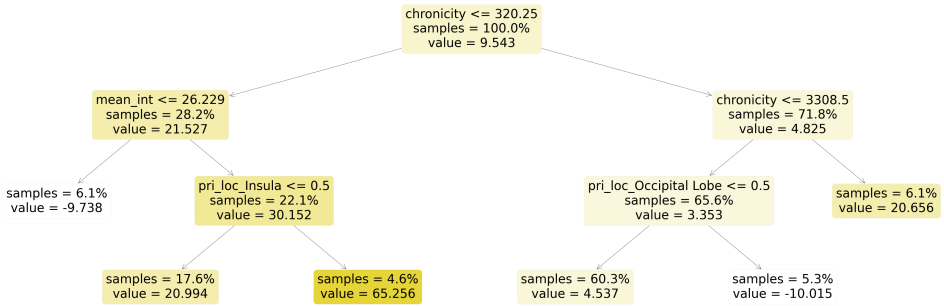
Figure 4.5: Decision trees predicting Dice score from metadata alone for baseline model. 'value' corresponds to the predicted Dice score for all samples lying at that node. Categorical metadata is one-hot encoded, thus 'label_name $\leq$ 0.5' means for the volumes on the right side split the label is present, and for the volumes on the left split it is not.

The main feature that gives good predictive performance is a primary lesion location of 'Temporal Lobe'. This is the second largest lobe in the brain after the frontal lobe, so it could be that there are many lesion examples in this area in the training set to learn from.

Fig. 4.6a shows a decision tree predicting Dice score for predictions made with a FiLMed model conditioned with Primary and Secondary Stroke Locations - the best performing model using non-derived metadata. It can be seen that the model performs particularly badly on cases imaged earlier than 19 days post stroke - this could be due to lesions across locations appearing similar in the early stages, and only having distinct appearances in later stages post stroke. The best performance is seen for lesions lying in both the Frontal and Temporal Lobes, however, as these are the two largest lobes in



(a) Locations FiLMed Model, Vol Removed: Dice Score



(b) Locations FiLMed Model, Vol Removed: Percentage Increase from Baseline Dice Score

Figure 4.6: Decision trees predicting Dice score (above) and percentage increase in baseline Dice score (below) from metadata alone for a FiLMed model conditioned with Primary and Secondary Stroke Locations.

the brain this could be seen as a proxy decision node for denoting strong performance on large lesions.

Fig. 4.6b goes further - instead of predicting the Dice score for the Location FiLMed model, the percentage increase in Dice score over the baseline model is taken as the predicted value. This shows clearly the metadata characteristics of the volumes where the lesion segmentation is *most* improved by incorporating metadata. Following the decision nodes down to the leaf with the highest percentage increase, we see the subset of volumes with chronicity under around a year, mean lesion intensity above around the average (28), and a primary stroke location in the Insula demonstrate the best improvement. These nodes show that adding stroke location as conditioning information, allows the network to learn the nonlinear trend observed in *Fig. 3.5* between chronicity and mean lesion intensity. The Insula is one of the smallest regions in the brain - this large increase in performance for the FiLMed model suggests these lesions have visual characteristics that are specific to the region that the baseline model fails to capture.

Chapter 5

Conclusion

The research conducted in this dissertation shows that image and patient metadata can offer an improvement in ischaemia segmentation when used to condition a U-Net with FiLM layers. This increase in performance can be up to 7.3% from baseline U-Net performance, which is achieved when incorporating a combination of readily available image and patient metadata, and metadata derived from ground truth labels. Discounting models trained with derived metadata, the best performing model is shown to give a 6.3% improvement by incorporating information about Primary and Secondary Stroke Locations. Incorporating information about stroke laterality is also shown to be effective, showing that spatial information can be useful to condition FiLMed networks with (in contrast to findings from previous research [17]). Although FiLM conditioning does not directly allow for spatial modulation, this is thought to be due to lesions in different anatomical regions presenting across volumes with consistently distinct visual features. Out of the two fusion methods investigated, late fusion is shown to have less variation in performance between runs with also good trade-off between training time and performance improvement. However, ultimately it is the models using complete fusion that show superior performance. The metadata prediction experiments show that if metadata information is already completely encoded in the images themselves, such as the scanner type, it can at best offer no improvement over baseline to condition a U-Net model with, and at worst results in a drop in performance. Lesion location however, is shown to be largely encoded within the images alone, yet still improves performance when used as FiLM conditioning input. As shown in *Fig. 4.6* where models with location label 'Insula' are predicted with very high accuracy, this is perhaps due to the metadata input offering large increases in performance in the cases where certain lesion locations are not already encoded in the images. We also find that this

method is most effective at improving very low Dice scores, motivating the use of this method in lesion segmentation where the consequences of failing to detect a lesion are severe.

5.1 Future Work

This research has some limitations that could be improved on in future work. The overall low Dice scores reflect the complexity of the task of automatic ischemic lesion segmentation, but could be slightly improved with more extensive hyperparameter tuning than this research timeline allowed for. Paing et al. [26] reach a Dice score of 0.6087 using a U-Net with the same dataset, although the total training time of this model is 42.2 hours. It is possible that with optimised hyperparameters for each different FiLMed model, metadata conditioning with FiLM layers would offer an even more significant increase in performance.

The original use case of FiLM layers is for visual question answering, where there is a very direct link between the conditioning information (question about an object in the image), and image itself. As shown in *Fig. 4.2*, for stroke ischaemia MRI, there is a varying relationship between metadata and the MRI images themselves for different types of metadata. Wolf et al. [27] propose a method that deals with this varying correlation between images and tabular metadata. The method is similar to FiLM, except rather than simply allowing the tabular metadata to modulate the network, baseline MRI features are concatenated with the tabular metadata to generate the FiLM parameters that layers are then modulated with. This allows for a bi-directional flow of information between metadata and image features, before feature map modulation occurs. The authors show this is successful for a disease classification problem, however a future research direction could be to extend this to the lesion segmentation task presented in this dissertation.

Bibliography

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[2] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," 2018.

[3] S.-L. Liew, J. M. Anglin, N. W. Banks, and M. Sondag, "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations.," *Sci Data*, vol. 5, 2018.

[4] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," 2017.

[5] V. L. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World stroke organization (WSO): Global stroke fact sheet 2022," *Int J Stroke*, vol. 17, pp. 18–29, Jan. 2022.

[6] J. M. Cassidy, G. Tran, E. B. Quinlan, and S. C. Cramer, "Neuroimaging identifies patients most likely to respond to a restorative stroke therapy," *Stroke*, vol. 49, Jan. 2018.

[7] W. W. Boonn and C. P. Langlotz, "Radiologist use of and perceived need for patient data access.," *Journal of digital imaging*, vol. 24, 2009.

[8] M. Person, M. Jensen, A. O. Smith, and H. Gutierrez, "Multimodal Fusion Object Detection System for Autonomous Vehicles," *Journal of Dynamic Systems, Measurement, and Control*, vol. 141, 05 2019.

[9] T. Trzcinski, "Multimodal social media video classification with deep neural networks," in *Photonics Applications in Astronomy, Communications, Industry,*

and High-Energy Physics Experiments 2018 (R. S. Romaniuk and M. Linczuk, eds.), vol. 10808, p. 108082U, International Society for Optics and Photonics, SPIE, 2018.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[11] T. Kim, I. Song, and Y. Bengio, "Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition," pp. 2411–2415, 08 2017.

[12] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2017.

[13] "Combining cnn-based histologic whole slide image analysis and patient data to improve skin cancer classification," *European Journal of Cancer*, vol. 149, pp. 94–101, 2021.

[14] C. Ou, S. Zhou, R. Yang, W. Jiang, H. He, W. Gan, W. Chen, X. Qin, W. Luo, X. Pi, and J. Li, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, 2022.

[15] A. Lemay, C. Gros, O. Vincent, Y. Liu, J. P. Cohen, and J. Cohen-Adad, "Benefits of linear conditioning with metadata for image segmentation," 2021.

[16] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101535, dec 2019.

[17] G. Jacenków, A. Q. O'Neil, B. Mohr, and S. A. Tsaftaris, "Inside: Steering spatial attention with non-imaging information in cnns," 2020.

[18] O. Vincent, C. Gros, J. P. Cohen, and J. Cohen-Adad, "Automatic segmentation of spinal multiple sclerosis lesions: How to generalize across mri contrasts?," 2020.

[19] I. Sheth, A. A. Rahman, M. Havaei, and S. E. Kahou, "Pitfalls of conditional batch normalization for contextual multi-modal learning," 2022.

[20] A. Pinto, R. Mckinley, V. Alves, R. Wiest, C. A. Silva, and M. Reyes, "Stroke lesion outcome prediction based on mri imaging combined with clinical information," *Frontiers in Neurology*, vol. 9, 2018.

[21] D. Kawahara and Y. Nagata, "T1-weighted and t2-weighted MRI image synthesis with convolutional generative adversarial networks," *Rep Pract Oncol Radiother*, vol. 26, pp. 35–42, Feb. 2021.

[22] G. Jacenków, A. Chartsias, B. Mohr, and S. A. Tsaftaris, "Conditioning convolutional segmentation architectures with non-imaging data," 2019.

[23] "U-net explained." https://towardsdatascience.com/u-net-explained-understanding-its-image-segmentation-architecture-56e4842e313a. Accessed: 2023-09-13.

[24] M. Nauta, R. van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," 2021.

[25] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting cnns via decision trees," 2019.

[26] M. P. Paing, S. Tungjitkusolmun, T. H. Bui, S. Visitsattapongse, and C. Pintavirooj, "Automated segmentation of infarct lesions in t1-weighted mri scans using variational mode decomposition and deep learning," *Sensors*, vol. 21, no. 6, 2021.

[27] S. Pölsterl, T. N. Wolf, and C. Wachinger, "Combining 3d image and tabular data via the dynamic affine feature map transform," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 688–698, Springer International Publishing, 2021.