Imperial College London

Individual Project Report

Imperial College London

Department of Computing

Tackling Crohn's disease using deep learning

Author:
Feifan Fan

Supervisor:
Dr. Bernhard Kainz
Second Marker:
Prof. Ben Glocker

Submitted in partial fulfillment of the requirements for the MEng Mathematics and Computer Science of Imperial College London

Abstract

This is the abstract about this thesis.

Acknowledgements

Thanks mum!

Contents

List of Figures

Chapter 1

Introduction

1.1 Crohn's Disease

Crohn's Disease [1, 2] is one of the primary types of Inflammatory Bowel Disease (IBD) [3], which is characterised by its chronic nature with inflammation in the gastrointestinal (GI) tract, as indicated in Figure 1.1 [4]. This condition can lead to long-term damage and complications, such as strictures, fistulas, and abscesses. Many people worldwide are struggling with IBD, and the management remains a challenge for medical professionals to address. A study from the University of Nottingham [5] reports that more than half a million individuals in the UK are affected by Crohn's Disease and Ulcerative Colitis, another significant IBD subtype. Unlike Ulcerative Colitis, which is limited to the colon and rectum, Crohn's disease can potentially develop lesions anywhere within the GI tract. Consequently, patients may experience diverse symptoms, including abdominal pain, diarrhoea, fatigue, and weight loss.

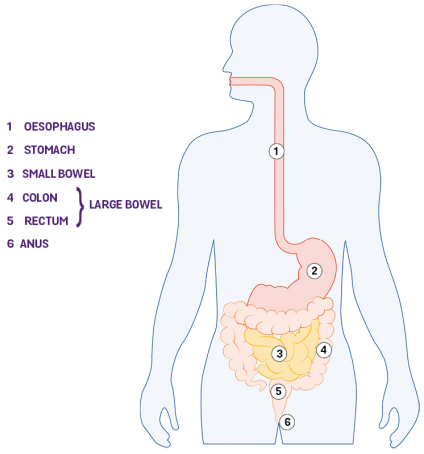


Figure 1.1: The gastrointestinal tract of a Human

Although numerous research initiatives have been undertaken [6, 7], the precise aetiology of Crohn's disease remains elusive, rendering it incurable. However, fortunately, early diagnosis and appropriate treatment can alleviate patients' symptoms and substantially improve their quality of life. Various diagnostic methods are employed by clinicians, such as enteroclysis, endoscopy, colonoscopy, and radiographic techniques (including barium contrast X-rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI)) to assist in the early diagnosis of the disease. MRI has become increasingly popular among radiographic techniques due to its non-invasive nature and enhanced imaging capabilities compared to CT. Nevertheless, manual MRI scan analysis remains a challenge since it is a time-consuming and labour-intensive process. Additionally, medical experts must examine each scan slice by slice painstakingly.

1.2 Motivation

The advancement of Machine Learning and Deep Learning technologies, notably Convolutional Neural Networks (CNNs), offers powerful means for automatic feature extraction from input imaging data, thereby supporting medical professionals in diagnostic tasks. One crucial aspect of Crohn's Disease diagnosis is the examination of the terminal ileum (T.I.). Holland et al.'s study [8] in 2019 proposed a residual network specifically targeting the terminal ileum to facilitate automated detection of Crohn's Disease using MRI scans. The authors claimed that the efficacy of their framework was contingent upon the degree of localisation during the preprocessing stage. Consequently, they advocated for incorporating terminal ileal ground-truth segmentations to enhance the localisation of the terminal ileum and improve the performance of automated detection techniques.

In a subsequent study, Abidi et al. [9] advanced this line of research by developing an innovative deep-learning tool based on the nnU-Net architecture [10]. This approach enabled the automatic localisation of critical regions, particularly the terminal ileum, essential for radiologists during diagnostic assessments. The researchers addressed the previously identified limitations regarding the high dependence on localisation during the preprocessing phase [8]. Furthermore, their findings established a solid foundation for a multi-class terminal ileum segmentation algorithm that combines transfer learning strategies with the nnU-Net architecture.

Inspired by the insights from [8, 9], this project aims to create a binary segmentation model capable of accentuating terminal ileum regions more accurately by incorporating advanced transfer and semi-supervised learning methodologies. The successful accomplishment of this objective will significantly aid clinicians in diagnosing Crohn's disease, ultimately contributing to enhanced patient outcomes.

1.3 Machine Learning Challenges

The efficacy of deep learning models is intrinsically linked to the training data's quality, quantity, and diversity. One of the principal challenges faced in this project is the limited availability of training data. The dataset at our disposal is relatively small, comprising only 233 patient cases, which pales compared to those utilised in other industry-leading deep learning systems. Furthermore, since the region of interest (ROI) occupies a minor portion of the MRI scan, additional preprocessing techniques, such as localisation and cropping, must be considered for enhancing the segmentation model's performance, as suggested by [9]. Another major challenge is the necessity for manual segmentation by clinical experts to develop gold-standard labels or point-wise centerlines for patient data, which is a complex, laborious, and inefficient endeavour. Consequently, acquiring high-quality and abundant patient data and gold-standard annotations poses significant challenges.

To mitigate these concerns, we propose a proxy training task employing weak supervision to generate coarse-grained segmentation masks as a compromise for the scarcity of gold-standard segmentations. Upon completion of the proxy task, gold-standard segmentations will be integrated into the training process to produce the final segmentation model. However, prior research [9] indicates that training from scratch using the nnU-Net framework for proxy tasks is inefficient due to long convergence times and unstable performance. Implementing transfer learning for proxy training [11] or incorporating a related pre-training job may serve as potential solutions to address these limitations.

1.4 Objectives

The primary of this project is to leverage deep learning methodologies to construct a progressive terminal ileum segmentation model. To achieve this, we will draw upon prior research whilst incorporating sophisticated transfer and semi-supervised learning techniques. The specific objectives that inform our strategic approach include:

- **Model Implementation**: We aim to establish a nnU-Net-based baseline segmentation model. This endeavour will necessitate the integration of appropriate data preprocessing techniques to optimise model performance.
- **Proxy Training Task Setup**: For data devoid of ground-truth annotations, we propose to generate coarse-grained weak masks based on precedents set by previous research. These masks serve as a foundation for establishing the proxy training task.

- **Target Model Development**: After training the proxy model using the generated weak masks, we plan to use the proxy model in conjunction with fully annotated data to develop our target segmentation model. This target model will set the baseline for subsequent refinement.

- **Model Refinement**: Once our baseline is set, we plan to enhance the generation of weak masks by integrating the SegmentAnything Model [12] from Meta AI. This will influence the training of a refined model, promising improved segmentation outcomes.

- **Performance Evaluation**: An essential part of our project is monitoring and quantitatively evaluating the performance of all developed models. We will compare the Dice Similarity Coefficient (DSC), scrutinise training efficiency, and assess the generalisation gap to ensure our models meet the established accuracy and efficiency standards.

Chapter 2

Background

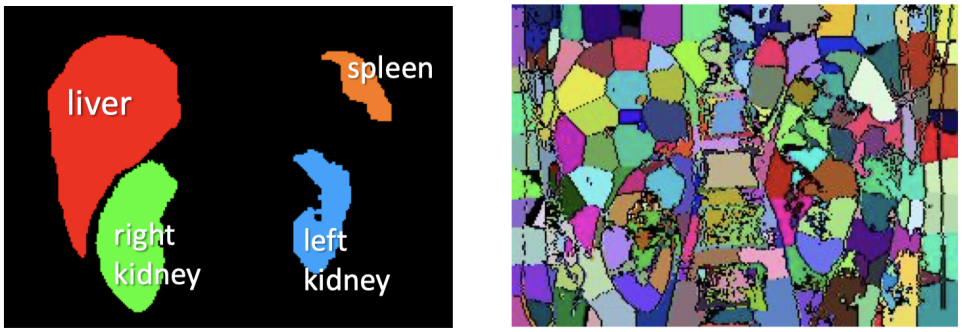
2.1 Delving into Image Segmentation

Image Segmentation is a fundamental concept in the realm of computer vision, which demonstrates how machines perceive and understand image data. But what exactly does the term "Image Segmentation" encapsulate? To explain this, let's take an illustrative example. Imagine a picture featuring two birds. The task of image segmentation, or more specifically, **"semantic segmentation"**, involves dissecting the entire image into distinct regions that are assigned different colour codes. These regions delineate the exact position of the birds within the image, effectively separating them from the background.

2.1.1 Expanding on Semantic Segmentation

Semantic segmentation transcends the act of partitioning an image into various regions. It assigns each segmented region with a label, i.e. a **semantic meaning**, indicating what the region denotes. An illustrative example is provided in Figure 2.1a, where the red region denotes the liver in a medical imaging scan. Returning to our bird image, we would assign these regions the label "bird" upon segmenting the regions corresponding to the birds. Similarly, the rest of the image, or the background region, would be labelled as "background".

One worth mentioning is that semantic segmentation distinguishes itself from other tasks that merely cluster images into coherent regions, as shown in Figure 2.1 [13]. The regions identified through semantic segmentation carry a specific value or meaning inherently linked to the task at hand. Simply put, not just 'where' but 'what' is just as crucial in semantic segmentation. Having laid out an overview of **Semantic Segmentation**, we will delve deeper into the distinctive methodologies utilised for segmentation tasks in the ensuing sections. This preliminary understanding provides a critical foundation for the exploration of more complex segmentation strategies and



(a) Semantic Segementation (b) Segmentation based on clustering

Figure 2.1: Different Image Segmentation Tasks

their applications in tasks such as diagnosing diseases or enhancing medical imaging methods.

2.2 Manual Segmentation

Manual Segmentation, while basic, forms a cornerstone in the realm of image segmentation. In the context of our project, it involves utilising the specialised knowledge of medical experts to meticulously create 'gold standard' labels for MRI data. This process is not simply a cursory glance at the image; it requires careful slice-by-slice examination of the MRI data, highlighting the diseased from the healthy tissues. This level of detailed inspection ensures that such manual segmentation results in validated and trustworthy classifications.

However, manual segmentation does bring to light significant challenges. Notably, the time-intensive nature of this process becomes increasingly pronounced when dealing with large datasets. This situation is further compounded in scenarios requiring slice-by-slice analysis of MRI data. The resultant effect is that manual segmentation can be exceptionally time-consuming, limiting the volume of data that can be labelled within an acceptable timeframe.

This leads us to confront an ever-present challenge: the scarcity of gold standard labelled data. The considerable investment of time and skilled resources required for manual segmentation naturally restricts the availability of such high-quality, expert-classified data sets. As we journey towards leveraging machine learning methodologies in the realm of medical imaging, this scarcity of gold-standard, manually segmented data surfaces as a substantial hurdle to overcome. Thus, innovative solutions are required to address this gap and enhance the effectiveness and efficiency of image segmentation processes.

Additionally, manual segmentation may introduce variability. This in-

cludes both inter-observer variability, where there could be disagreement between different human observers, and intra-observer variability, where inconsistencies could arise from the same observer at different occasions. This variability might potentially compromise the validity of the manually segmented data.

The process of manual segmentation may also unintentionally incorporate biases from human experts, which could potentially distort the results. The limitations inherent in manual approaches underscore the necessity for the integration of computational tools to bolster the segmentation process. By leveraging computer-assisted methodologies, we can introduce an intelligent strategy to carry out segmentation tasks. This approach not only enhances efficiency but also promotes consistency, thereby effectively minimising human error and variability. As a consequence, we can significantly enhance the overall reliability and accuracy of MRI data analysis. This confluence of human expertise and computational intelligence stands to advance the field of medical imaging analysis.

2.3 Threshold-based Methods

Thresholding is among the most rudimentary yet widely adopted methods for image segmentation in various industry applications. The premise of this technique rests on the hypothesis that the distribution of pixel intensity in an image contains multiple modes. In other words, the grayscale intensity of pixels can be differentiated into two (or more) distinct clusters. The strategy involves identifying thresholds within the 'gaps' separating these clusters to delineate the background from the foreground. For accomplishing this segmentation, pixels exhibiting an intensity lower than the threshold are designated as background, whilst those with equal to or higher than the threshold are identified as foreground. Furthermore, it is also feasible to define multiple thresholds, thereby enabling multi-class segmentation of pixels that fall within a specific intensity range only.

The inherent simplicity and speed of this algorithm is undeniably advantageous. However, its applicability is contingent on the homogeneity and distinctness of the regions of interest (ROIs), implying that pixels within the same region should exhibit similar intensity values. An additional challenge lies in identifying consistent threshold values across different images. This is due to the fact that pixel intensities are prone to variations between different images, thus introducing a degree of complexity in maintaining uniform thresholding norms across disparate datasets.

2.4 Region-based Methods

Region-based segmentation techniques, with the region-growing method standing as a prominent example, offer an alternative approach to image segmentation. This method hinges on the premise of homogeneity within the segmented regions. The algorithm initiates with a seed pixel and then expands the region by successively incorporating pixels similar to the initial seed. The process continues until the region growth reaches a pre-defined size or once the region achieves homogeneity - interpreted as the point when neighbouring pixels become significantly dissimilar [14].

Utilisation of the region-growing method often proves efficient, generating a connected region starting from the seed point. Unlike thresholding methods that rely on explicit image properties, region-growing methods facilitate segmentation based on pixel similarity. However, this method exhibits sensitivity to noise, as the algorithm may persist in growing the region even if the neighbouring pixels significantly deviate from the seed pixel's properties.

Another key challenge when employing the region-growing method is the critical significance of the initial seed point selection. An incorrect choice of the seed point could thwart proper region growth. This initial phase often becomes time-consuming and lacks accuracy when seeking the optimal seed point. Furthermore, it necessitates human intervention for evaluating the appropriateness of the chosen seed point.

To Address these concerns in the context of medical imaging, a study contucted by Poonguzhali et al. proposed an automated region-growing method tailored for ultrasound images [15]. Their approach introduces an automatic seed point selection mechanism predicated on textural features, such as co-occurrence and run-length features, thereby eliminating the need for manual intervention. The experimental results attested to the feasibility and efficacy of their proposed region-growing algorithm, showcasing its capacity to select seed points and segment the ROIs without requiring manual intervention.

2.5 Deep Learning Methods

2.5.1 U-Net: An Automated Deep Learning Method for Image Segmentation

U-Net [16] represents a transformative approach in the realm of deep learning methodologies for image segmentation, particularly aligned with our project's objectives. As an evolved variant of Fully Convolutional Networks (FCN) [17], U-Net integrates a high-level contextual extraction path with a symmetric localisation pathway. This unique arrangement both captures a broad context of the image and enables precise localisation. This is ideal for creating an automated pipeline for T.I. segmentations, and an essential requirement in our research.

The end-to-end learning facilitated by U-Net directly generates pixel-wise segmentation masks from raw pixels, creating a valuable asset for our project. The architecture of this robust network is illustrated in Figure 2.2.

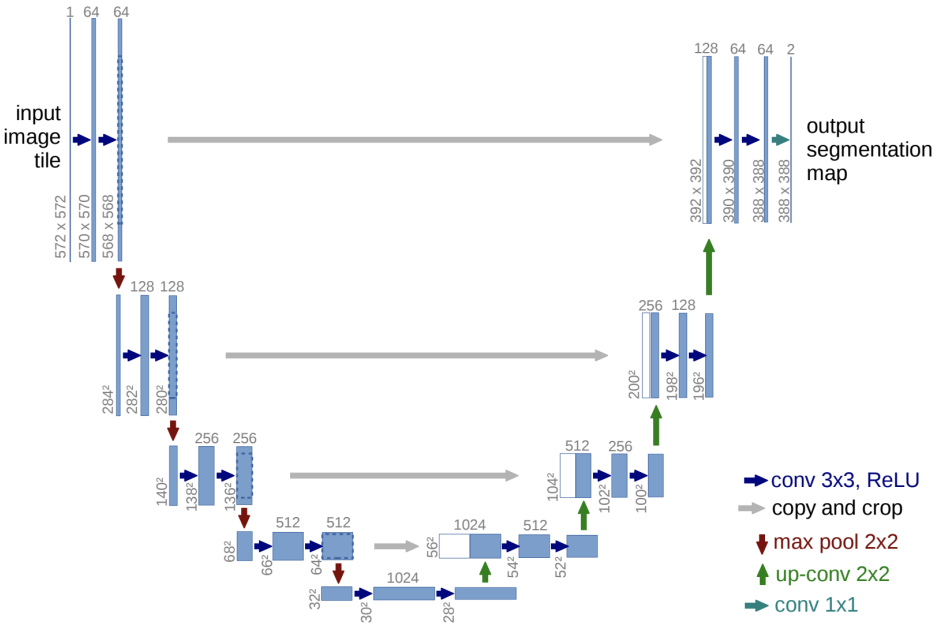


Figure 2.2: An example of the U-net architecture.

Significantly, U-net has also been developed to handle 3D data, which aligns perfectly with our scenario, where the MRI scans are inherently 3-dimensional images. Furthermore, each blue box in Figure 2.2 encapsulates a multi-channel feature map. The number of channels is denoted at the top of each box, while the x-y dimensions are annotated at the lower left corner of the box. Meanwhile, the white boxes signify sets of copied feature maps, with arrows demonstrating varied operations.

On the contracting (left) side of the network, the design entails a series of recurrent steps. Each step initiates with two 3x3 convolutions, proceeding to double the number of feature maps. The result is then subjected to a Rectified Linear Unit (ReLU) activation function, followed by a downsampling operation using 2x2 max pooling with a stride of 2.

The contracting path of the network (left) contains recurrent steps, each initiating with two 3x3 convolutions that effectively double the feature maps count. Subsequently, a Rectified Linear Unit (ReLU) activation function is applied, followed by a 2x2 max pooling downsampling operation with a stride of 2.

On the expansive side (right) of the network, the process commences with an upsampling step. A 2x2 convolution follows, halving the feature

channels, which are then concatenated with their corresponding features from the contracting pathway. To conclude, a pair of 3x3 convolutions are applied to the image, succeeded by ReLU activations. The final layer employs a 1x1 convolution that maps each 64-component feature vector to the desired classes.

Overall, U-Net offers a comprehensive and highly effective tool for achieving granular image segmentation—a capability integral to the success of our research exploration.

2.5.2 nnU-Net: A Self-configuring Framework Aligned with Our Research Goals

The nnU-Net [10] offers a progressive step towards personalised segmentation techniques. As a framework built upon U-Nets, it embodies a self-configuring segmentation mechanism, which autonomously orchestrates the configuration of preprocessing, network architecture, training, and post-processing steps in a segmentation pipeline. Crucially, the configuration selected is not static but instead adapts to the specificities of the medical data used for training.

A standout feature of nnU-Net is its unique approach towards determining hyperparameters. The framework utilises "data-fingerprints" allied with heuristic rules to pinpoint the optimal hyperparameter configuration for a given dataset prior to processing the training data. This data fingerprint concept is further leveraged to generate pipeline configurations, encapsulating both inferred parameters (such as image resampling, normalisation, batch, and patch size) and blueprint parameters (such as loss function, optimiser, and network architecture).

With these pre-selected hyperparameters and generated pipeline configurations, nnU-Net proceeds to facilitate network training for 2D, 3D full-resolution, and 3D-Cascade U-Nets [18]. The platform then select an ensemble of configurations from these three networks to achieve optimal performance (for instance, maximising the average dice coefficient). Once identified, this optimal configuration is subsequently deployed and evaluated on the test dataset.

Notably, nnU-Net has been shown to deliver state-of-the-art performance in various medical imaging tasks, including segmenting brain tumours, prostate tissues, and liver structures [19].

However, it is essential to understand that the 'adaptive' nature of nnU-Net does not imply universal applicability. Once the framework is trained, it performs optimally when applied to new data similar to the training data. For instance, if nnU-Net is trained using abdominal MRI images for liver segmentation, it will yield the best performance in similar tasks involving abdominal MRI images.

Therefore, while nnU-Net enables a highly adaptive and automated seg-

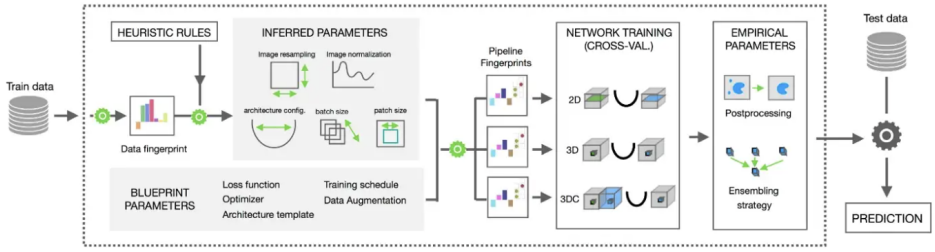


Figure 2.3: The pipeline representation of nnU-Net.

mentation workflow, the utility of the trained model is maximised when deployed on tasks that closely resemble the context and content of the training data. In our specific case, we will train the nnU-net model with our specific type of 3D MRI scans for T.I. segmentation tasks to ensure optimal results. This approach ensures leveraging the full potential of nnU-Net's adaptive capabilities tailored to our research goals.

As our research strives to establish an automated pipeline for T.I. segmentations on 3D MRI scans, the integration of the nnU-Net framework is poignantly relevant. With its inherent capacity to adapt to different datasets and consistently yield precise image segmentation, nnU-Net is excellently positioned to enhance the rigour and precision of our research endeavour.

2.6 Generating Weak Labels: A Strategic Response to Annotation Scarcity

Given the scaricity of gold standard annotations for segmentation tasks, as discussed in section 2.2, it becomes imperative to identify effective strategies that can enhance our training outcomes. One potential approach is the use of weak labels or weakly-segmented masks derived from the data. By adopting such methods, we aim to extract critical image features that can contribute to the learning efficacy of our model, despite limited access to manually-annotated, gold standard data.

2.6.1 Leveraging Unsupervised Methods: Simple Linear Iterative Clustering (SLIC)

A notable unsupervised method we contemplate employing is the Simple Linear Iterative Clustering (SLIC) [20]. This algorithm stands at the forefront of superpixel segmentation techniques, providing a powerful tool to cluster pixels within an image into compact and uniformly labelled regions, referred to as 'superpixels'.

The key strengths of SLIC are its simplicity, ease of implementation, and adaptability across diverse scenarios. These virtues enable SLIC to effectively handle boundary adherence issues while simultaneously reducing the computational burden associated with image segmentation tasks. Essentially, SLIC operates as a spatially constrained iterative k-means clustering method with a pre-determined superpixel size m. Consequently, in the generated weak mask, each superpixel generated incorporates a cluster center and m pixels. The SLIC algorithm adapted from [20] is illustrated in Algorithm 1:

Algorithm 1 SLIC superpixel segmentation

1: Initialise cluster centers $C_k = [l_k, a_k, b_k, x_k, y_k]^\top$ by sampling pixels at regular grid with grid intervel S.
2: Perturb cluster centers in a $n \times n$ neighbourhood, to the loweset gradient position.
3: **while** $E >$ threshold **do**
4: **for** each cluster center C_k **do**
5: Assign the best matching pixels from a $2S \times 2S$ square neighbourhood around the cluster according to the distance measure [20].
6: **end for**
7: Compute new cluster centers and residual error E.
8: **end while**
9: Enforce Connectivity.

Note the term "Perturb" in line 2 essentially means to re-assign the cluster center in a $n \times n$ neighbourhood, where the re-assigned center is the point has lowest gradient position compare to the original cluster. Additionally, the residual error E is defined as the L_1 distance between previous centers and recomputed centers.

By integrating SLIC as part of our weak label generation strategy, we aim to utilise these inherent benefits to enhance the robustness of our machine learning models, thereby bringing rich insights from our medical imaging data in the absence of extensive manually-curated annotations.

2.6.2 Harnessing Pretrained Models: The Segment Anything Model (SAM)

Emerging from the innovative approaches of Meta AI, the Segment Anything (SA) project presents a new paradigm for image segmentation. The cornerstone of this project is a pioneering model, known as SAM, which exhibits an impressive degree of adaptability and transferability. Specifically designed and trained to be promptable, SAM showcases remarkable capabilities in zero-shot transfer to new image distributions and tasks.

This inherent agility of SAM instigates impressive zero-shot performance outcomes. Indeed, comparative analysis with prior fully supervised results reveals that SAM often delivers competitive, if not superior, performance metrics [12].

Extending this versatile approach to medical imaging, MedSAM offers compelling prospects. Derived from the parent SAM model, MedSAM has shown considerable promise in segmenting medical images. One study unearthed that SAM's functionality can be enhanced with manual prompts, such as points and boxes, indicating intended objects in medical images [21]. This enhancement strategy aligns perfectly with our proposed approach to initially employ SLIC for coarse segmentation, and subsequently refine the segmentation process by obtaining finer-grained weak masks from SAM.

However, it's worth noting that whilst SAM exhibits remarkable performance with certain objects and modalities, it may fall short or even fail in other scenarios. This observation underscores the importance of particularising the model's utility in context-specific applications, a perspective central to our research objective on T.I. segmentations with 3D MRI scans. With its inspiring capabilities, SAM holds immense potential to enhance our quest for automated and precision-driven segmentation techniques, thereby bolstering our research outcomes.

Chapter 3

Related Work

3.1 Automatic Detection and Segmentation of Crohn's Disease Tissues from Abdominal MRI

The origins of applying deep learning techniques to medical image analysis can be traced back to approximately a decade ago. In 2013, Mahapara et al. [22] pioneered a machine learning-based method for segmenting bowel regions to detect Crohn's disease tissues in MRI scans.

The proposed pipeline begins with the over-segmentation of the input MR image test volume into supervoxels. Sequentially, Random Forest (RF) classifiers are employed to identify supervoxels containing diseased tissues, subsequently defining the Volume of Interest (VOI). Within the VOI, voxels are further examined to segment the affected region. An additional set of RF classifiers is applied to the test volume to generate a probability map, which delineates the likelihood of each voxel being classified as diseased tissue, normal tissue, or background.

Empirical results demonstrate that this approach achieved satisfactory segmentation performance, as evidenced by a Dice metric value of 0.90 ± 0.04 and a Hausdorff distance of 7.3 ± 0.8 mm. The significance of this research lies in the development of an automated pipeline for segmenting diseased bowel sections in abdominal MR images. This pipeline assists medical experts in identifying affected tissues, thereby facilitating the diagnosis and treatment of Crohn's Disease. The clinical validation of the results, showcasing high segmentation accuracy, further underscores its utility in supporting medical professionals in their work.

Nevertheless, the method is limited by computational inefficiency and complexity due to extended testing times for each instance, fine-tuning requirements for each pipeline stage, and ample opportunities for architectural improvements. Moreover, the research does not delve into finer details, such as the terminal ileum, which is particularly crucial for comprehensive analysis and early diagnosis.

3.2 Automatic Detection of Bowel Disease with Residual Networks

Building on several years of research in the field, Holland et al. [8] put forth a pioneering approach in 2019 to automate the detection of Crohn's disease from a limited dataset of MRI scans. The authors employed an end-to-end residual network [23], equipped by a soft attention layer [24]. This layer essentially magnified salient local features and added a layer of interpretability, providing a clearer understanding of the analytical process.

In a strategic departure from semantic segmentation strategies typically employed, their approach exclusively targets the terminal ileum. This focus served to underscore the potential feasibility of deep learning algorithms for the precise identification of terminal ileum Crohn's Disease within abdominal MRI scans.

The method's robustness is reflected in its experimental results. Under conditions of localized data within a semi-automatic setting, the model achieved a commendable weighted-f1 score of 0.83. This score is particularly noteworthy given its close correlation with the MaRIA [25] score, a clinical standard that enjoys widespread acceptance in the medical community. Beyond its performance metrics, the researchers accentuated the relative efficiency of their model, which necessitated only a fraction of the preparation and inference time compared to standard procedures. This aspect underlines the potential for significant time-saving benefits in a clinical context.

However, the research did reveal certain limitations. Notably, when applied in a fully automatic setting, the model's performance exhibited a marginal decrease in efficiency. Although this does not detract from the overall achievements of the study, it does highlight an area where further refinement and improvement could be pursued.

Reiterating their discovories, Holland et al. proposed a strong correlation between model performance and the degree of localisation in the training data. They suggested the collection of gold-standard segmentation of the terminal ileum could prove beneficial as an antecedent task in efforts to enhance automatic detection performance. This proposition opens up intriguing possibilities for research, including the work presented in this thesis, which explores these aspects in greater detail.

Their insights illuminate the synergistic potential between manual analysis and automated methods in enhancing diagnostic capabilities. Importantly, they establish a pathway for integrating deep learning techniques to detect Crohn's disease from limited datasets, indicating a promising approach to tackle one of the significant challenges in machine learning: data scarcity. By leveraging soft attention mechanisms to intensify salient local features and augment interpretability, they provide a valuable tool for medical professionals to comprehend better the results generated by the al-

gorithm.

These findings, particularly the proposed use of gold-standard segmentation of the terminal ileum, provide a solid foundation for the work pursued in this thesis.

3.3 Leveraging Machine Learning Methods for Accurate Prediction of Intestinal Damage in Crohn's Disease Patients

In 2020, Enchakalody et al. [26] embarked on an innovative study exploring the potential of machine learning methodologies to enhance the precision and reliability of diagnosing and monitoring Crohn's Disease. They applied these techniques to a small dataset of 207 CT-Enterography (CTE) scans, an approach that mirrors our own research focus. Their comprehensive analysis involved the intricate examination of cross-sectional views of small intestine segments and detailed detection of diseased tissues. Utilising two distinct classifier types - Random Forest (RF) with ensemble techniques and Convolutional Neural Network (CNN) algorithms, they quantitatively evaluated intestinal damage related to Crohn's Disease on each mini-segments.

The efficacy of both RF and CNN techniques was compellingly demonstrated in the experimental results, achieving accuracy rates of 96.3% and 90.7%, respectively, for classifying diseased and normal segments. Remarkably, these techniques mirrored the effectiveness of expert radiologists in distinguishing between diseased and normal small bowel tissue. This underscores the immense potential of machine learning, even when applied to small datasets, in elevating the precision of Crohn's Disease diagnoses.

The research conducted by Enchakalody et al. is particularly insightful for our work. It not only demonstrates the successful application of deep learning techniques on small datasets but also opens the door to potentially revolutionising the diagnosis and treatment of Crohn's disease through machine learning. It highlights the possibility of a more precise and automated approach to detecting intestinal damage in such patients, a focus that aligns closely with our current research aims.

While it should be noted that this study primarily focused on data derived from CT-Enterography, differing slightly from our focus on MRI data, the methodology and findings offer valuable insights. As of the time of writing this report, despite the progress made, achieving a fully automatic approach for diagnosing Crohn's disease based on cross-sectional imaging that equates to the proficiency of expert radiologists continues to be an exciting area of ongoing research.

Chapter 4

Ethical Discussion

Given the sensitive nature of our project that involves the handling of medical data, we are steadfastly committed to ensuring a robust ethical framework guides all phases of our work. This initiative encompasses further processing and augmentation of previously collected medical data, as well as merging existing datasets.

To safeguard personal identity and ensure strict adherence to privacy standards, all MRI scans are carefully processed under the oversight of clinical radiologists at St Mark Hospital. Comprehensive measures are employed to remove any personally identifiable data, such as names, genders, ages, and ID numbers, from the medical records. As a result, it is impossible to trace back any individual's identity from the processed medical data.

Our commitment to ethical considerations extends into the model training procedure. We employ the nnU-Net framework, which utilises convolutional layers to learn from data via feature extraction. This learning procedure does not store original training data; instead, it creates feature maps that represent distilled, valuable insights from the data. This process ensures that the original training data cannot be recovered from the model, thereby preserving individual privacy.

In essence, our project stands on a foundational ethical commitment that respects personal anonymity and data confidentiality. Recognising the crucial importance of trust in scientific inquiry, particularly when dealing with sensitive medical data, we are dedicated to exemplifying conscientious practices that uphold the highest standards of research ethics.

Chapter 5

Dataset Analysis

The focus of this thesis is the segmentation of medical images, specifically Magnetic Resonance (MR) images, with manually annotated gold standard segmentations of the colon and terminal ileum. MR imaging offers a rich depth of detail, attributable to its three-dimensional nature. This complexity, while advantageous for diagnosis and treatment planning, presents a unique set of challenges for image segmentation. It is imminently pertinent that the orientation in which these images are captured significantly impacts their interpretability and subsequent processing. In this chapter, we will delve into how our imaging data are obtained and perform exploratory analysis on the dataset.

5.1 Data Acquisition and Classification

At the heart of our research lies the utilisation of T2-weighted images, a type of MRI scan that provides detailed pictures of the inside of the body, to help with segmentation tasks. T2-weighted images are a specific genre of magnetic resonance imaging (MRI) scans manifested by the heightened intensity in fluid-rich structures, while their fat-laden counterparts appear darker. This distinctive contrast owes its origin to the heterogeneous responses of different tissues to the magnetic field and radio waves deployed during the MRI procedure.

The unparalleled detailing offered by T2-weighted images of internal human structures, especially soft tissues, makes them an invaluable asset in visualizing bodily fluids, identifying edema or swelling, and surfacing lesions - a potential indicator of Crohn's disease if detected in the gastrointestinal tract. Furthermore, they have been instrumental in diagnosing and assessing critical diseases, such as cancer and multiple sclerosis.

In enriching the diversity of our dataset, we have judiciously incorporated three variants of MR images. A summary describing the unique features and applications of each variant is as follows:

- **Axial T2-weighted Images**: Projected along the axial plane, these images provide a top-down representation of anatomical structures and are especially beneficial in examining anatomical correlations.
- **Coronal T2-weighted Images**: Imaged along the coronal plane, these scans offer a frontal perspective of the anatomy. These images are reputable for their proficient highlighting of fluid-filled structures and lesions.
- **Post-Contrast Axial T2 Images**: These images are captured after administering a contrast agent and enhance tissue contrast, thereby providing enriched insights into the nature of lesions.

The strategic incorporation of three distinct forms of MR images enriches our dataset, transforming it into a comprehensive repository designed to master the intricate task of segmentation. Yet, it is not devoid of potential obstacles such as MRI artefacts originating from patient movement or the overarching issue of data scarcity.

The limited availability of annotated data poses dual challenges—it affects not only the trajectory of the model training but also curtails our ability to effectively evaluate the model performance due to the restricted availability of testing data.

However, it is precisely this scarcity of data that fuels our quest for alternative annotated datasets to augment model training. The following chapter provides a detailed discussion of our innovative approach to addressing this challenge.

5.2 Dataset Specification

Our research utilises a carefully compiled dataset comprising 233 MR Images per class. This dataset balances representation with 113 abnormal cases and 120 normal cases for each image type. In addition to images, we also have a collection of centerline coordinates representing the colon in the MR image, although these are not available for every image. Furthermore, a select set of human-annotated ground truth results is included.

Upon closer examination, the centerline and ground truth annotation distribution is as follows:

The generation of colon centerline coordinates is a meticulous process, performed by radiologists via manual slice-by-slice inspection. Following their careful examination, they annotate the relevant slices with reference points linked to the colon. These annotations are archived as compressed XML files, colloquially referred to as **traces files**.

This storage format serves a dual purpose. Firstly, it facilitates visual inspection and analysis within the framework of medical imaging by clinical

Type of Image	Total Centerlines	Centerlines (abnormal:normal)	Ground Truth (abnormal:normal)
Axial T2	103	59:44	18:20
Coronal T2	93	46:47	18:30
Post-Contrast Axial T2	-	-	13:20

Table 5.1: Distribution of centerlines and ground truth segmentations for different MR images

experts. Secondly, it empowers developers to navigate through the XML tree to gather valuable information about the centerline coordinates. It is worth noting that the preliminary 20% of the points are often deemed the most accurate, typically representing the interval in which the terminal ileum is situated.

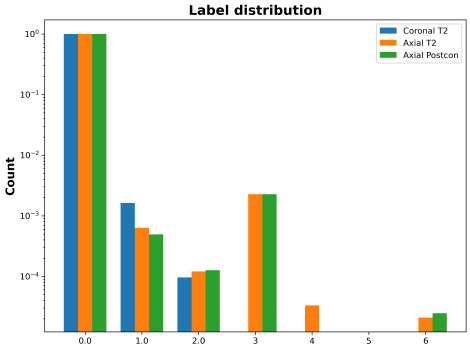
Our diverse and comprehensive dataset provides thoughtful insights for analysis and a solid foundation for reliable model training.

5.3 Ground Truth Segmentations

The ground truth segmentations delivered by our panel of clinical experts are assigned with varying semantic meanings, each corresponding to unique label IDs. The specifics of these relationships are highlighted in Table 5.1a. To gain a clear overview of the segmentation, I have extracted a subset of ten samples from each type of image data and performed manual segmentation. The ensuing label distribution, represented on a logarithmic scale, is graphically demonstrated in Figure Figure 5.1b.

Label	ID
Background	0
Abnormal T.I.	1
Normal T.I.	2
Colon	3
Colon	4
Appendix	6

(a) Label-ID representations used across manual segmentations



(b) Illustration of Segmentation Label Distribution

Figure 5.1: Detailed breakdown of segmentation specifics and label distribution across diverse categories of image data

From the results, a salient observation is the predominance of the background label in the voxel classes of the ground truth segmentation results. Its pronounced presence approximates to one, rendering other voxel labels relatively insignificant and emphasizing the challenges intrinsic to abdominal MRI segmentation.

This raised a need for us to rethink the segmentation scope. Given the relative insignificance of portions other than the background, the informational value they contribute toward the training process is marginal at best. This limited contribution not only impedes the ability of model to classify effectively within these regions but increases the risk of false-positive classifications considering the comparison between the remaining regions versus the background. However, to alleviate this issue, we simplify the scenario into a binary classification problem, interpreting all non-background voxels as general T.I. regions. Accurate segmentation of this region holds significant implications for enabling clinical experts to diagnose Crohn's disease at an early stage.

This insight has inspired us to consider utilizing the centerline coordinates as a guiding factor for future initiatives. Specifically, we are contemplating two potential applications: augmenting the data through bounding box applications or minimizing the search space by cropping the image. This could balance the label distribution between background and T.I. regions, thereby improving the suitability of the segmented images for subsequent analyses. Ultimately, these initiatives aim to overcome the challenges identified in our exploration of ground truth segmentations.

Chapter 6

Methodology

The training strategy for our segmentation model is composed of four pivotal stages: dataset preprocessing, weak label generation through proxy learning, fine-tuning with gold standard annotations in target learning, and ensemble learning. A detailed elaboration of these stages follows:

6.1 Dataset Preprocessing

Our exploratory data analysis unveiled that non-background voxels constitute only a minor proportion of the overall MRI volume. Consequently, they might not impart sufficient information to enable effective learning by the segmentation model. To rectify this issue, we exploit the provided centreline coordinates of the terminal ileum to define the bounding box encompassing the T.I. region. The original volume is then cropped to this reduced version for implementing the baseline, and the bounding box information is introduced for refining the weak labels with the pre-trained SAM model, which significantly curtailing data complexity and ensuring that the retained labelled voxels play a crucial role in facilitating model learning.

Prior to initiating training, the dataset must be restructured into a specific format as stipulated by nnU-Net [10]. As per this convention, the name of each dataset should adopt the form `'Dataset<ID>_<NAME>'`. Moreover, each dataset is accompanied by a configuration file providing essential metadata, such as the modalities of MRI imaging employed for each patient, along with the semantic interpretation of the segmented labels within the ground truth results. The images and corresponding labels assigned for training are situated within the dataset folder, separated into distinct directories: `'imagesTr'` and `'labelsTr'`. Note that we treated coronal and axial T2 as two separate datasets to complete the segmentation task.

Upon successful conversion of the dataset, nnU-Net initiates preprocessing, modifying images to different dimensions and resolutions optimised for ensemble model training. Concurrently, a unique **'dataset fingerprint'**

is derived, earmarked for subsequent hyperparameter tuning. With these preparatory steps completed, the preprocessed data is primed for the ensuing proxy learning task.

By adopting such a meticulous approach to data preprocessing, we ensure that our model is furnished with optimally structured data, thereby enhancing its capacity to learn effectively from the available information and generate accurate segmentations.

6.2 Baseline Implementation

Building upon precedent research, we've formulated a baseline model that adheres to a systematic and robust methodology. This model lays the groundwork for further enhancement and modifications later in our project:

Our first step involves the application of 3D Simple Linear Iterative Clustering (SLIC) superpixel segmentation on the cropped Region of Interest (ROI). This task serves to generate a weak mask that becomes instrumental in our proxy learning stage. To ensure the successful execution of this step, we identify the superpixels situated on the centreline of the terminal ileum, leveraging provided centreline coordinates. Subsequently, the segmented superpixels are relabelled to binary labels, as it realigns output with our project objectives.

Upon generation of the initial segmentation, we confront the potential issue of hole or tube-like structures within the segmented supervoxels. These irregularities can disrupt the continuity of the terminal ileum representation. Therefore, our process incorporates the use of a voting iterative binary hole filling algorithm in conjunction with morphological hole closing to create a more refined segmentation.

With the generation of weak masks accomplished, we proceed to train our proxy model. Furthermore, a second iteration of training is conducted, this time deploying the proxy model on fully-annotated data to obtain the final segmentation model. Additional aspects of the training process, along with the ensembling strategy employed to derive the final model, will be elaborated in the ensuing sections.

6.3 Model Training

As we step into our process, our attention is directed towards the pivotal stage of model training. This stage is characterised by an enriched proxy learning strategy, where we replace the initial unsupervised 3D SLIC-based weak mask generation with a more sophisticated tool: MedSAM. Additionally, the training and evaluation phases are designed to optimise model performance, incorporating cross-validation and ensemble learning methods. Below, we elaborate on these key aspects:

- **MedSAM Application**: In this subsection, we discuss in detail how we leverage MedSAM for generating high-quality weak masks, elucidating its superior capabilities in the context of medical image segmentation and its specific role in enhancing our model learning trajectory.

- **Dataset Partitioning**: Here, we explain how we divide our dataset into training and testing sets using an 80/20 ratio. We provide rationale for this split and discuss how it ensures a robust and unbiased evaluation of our model performance.

- **Training Pipeline**: Our multifaceted training pipeline starts with a two-phased approach. Initially, we train a proxy model on weak masks generated by MedSAM. This is followed by further training on 80% of the fully annotated data, maximizing learning from both weakly and fully annotated data sets.

  To ensure model robustness, a 5-fold cross-validation method is integrated during training to mitigate overfitting risks. Furthermore, our pipeline supports different configurations, which includes 2D and 3D variants of data, loss function adjustments, and variations in training optimisers and learning rates.

  The final aspect of our pipeline is ensemble learning, utilized to aggregate models across different folds and configurations, thereby generating our target model. The collective operation of these facets is meticulously designed to bolster our model performance in diverse scenarios.

- **Inference**: In this final stage, we delve into the utilization of the trained model for making inferences. Our focus lies in outlining the procedural aspects of applying the model to new and unseen data, thus generating predictive outcomes that form the basis for the subsequent comprehensive evaluation of our model performance. The evaluation process, with a detailed quantitative analysis, will be discussed extensively in the following chapter.

Through this comprehensively planned model training phase, we aim to build a proficient terminal ileum segmentation model that not only learns effectively from our data but also showcases robust performance across varying scenarios.

6.3.1 Refining Weak Masks with MedSAM

Enhancing the quality of our weak masks forms an integral part of our methodology. For this, we leverage a modern mask generation technique - MedSAM, a variant of the SegmentAnything Model. Being tailored for

medical images, MedSAM enables the creation of refined, granular weak masks, thereby setting the stage for improved segmentation results.

The process unfolds with our preprocessed images being introduced to MedSAM. Traditional methods pale in comparison to this sophisticated tool which, equipped with cutting-edge AI algorithms, excels in discerning the distinct characteristics intrinsic to medical images. The outcome is a granular segmentation that far surpasses the precision achievable with conventional models. A noteworthy augmentation to our preprocessing phase includes the inclusion of the extracted ROI along with the bounding box of centreline coordinates specific to the T.I. region. Evidence from Huang et al., 2023 [21], suggests that such enrichment significantly amplifies the quality of the resulting weak mask. For images lacking provided centerline information, we elect to exclude them from the dataset, thus ensuring a consistent and reliable source of data for our methodology.

Once these refined weak masks are generated, we align the mask labels to the binary format, thereby syncing it with our project objectives. Subsequently, these weak masks are employed during the proxy learning phase of our training pipeline.

The introduction of MedSAM into our pipeline marks a pivotal step towards superior segmentation outcomes - the refined masks open up enhanced learning avenues for our model, laying a solid foundation for further advancements in terminal ileum segmentation.

6.3.2 Executing Dataset Partitioning

The partitioning of our dataset into training and testing sets plays a crucial role in our model development. This strategy helps ensure that the model performance is assessed on new, unseen data, establishing a reliable measure of its predictive capabilities.

Our approach involves a partitioning of the dataset into an 80/20 ratio. Prior to this division, we shuffle the dataset to randomise the order of samples, ensuring an unbiased representation in both the training and testing sets. Subsequently, the first 80% of these randomized samples form our training set while the remaining 20% are reserved for testing.

The reason of using such ratio is because it is widely revered in the field of machine learning, serves to strike a balance between providing sufficient data to nourish the model's learning journey and withholding a substantial portion for an authentic evaluation of its performance. This balance becomes particularly crutial when navigating through scenarios characterised by a limited dataset, such as ours.

This cautious allocation of data equips us with a model that's not only well-trained but also rigorously tested for its predictive performance, thereby strengthening our confidence in its segmentation ability.

6.3.3 Establishing the Training Pipeline

This subsection outlines on principal components of our training pipeline, including the deliberate choices and reasonings behind formulating the loss function, selecting the optimiser and determining the learning rate. Additionally, a two-phased training approach is proposed with the key components collevtively to ensure the model's capability to learn efficiently, navigate an optimised prediction path and deliver robust T.I. segmentation.

Loss Function Formulation

Guiding the learning trajectory of our model during training is an elegantly designed loss function. Particularly, we harness nnU-Net's unique blend of cross-entropy and Dice losses [27], expressed as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Dice}$$

For our specific use-case, this composite loss function manifests into a binary form, simplifying to a binary cross-entropy loss defined as:

$$\mathcal{L}_{CE} = \sum_{i}^{N} \left(y_i \log o_i + (1 - y_i) \log(1 - o_i) \right)$$

In this equation, $y_i \in \{0, 1\}$ symbolizes the ground truth label for the i-th instance, while $o_i \in [0, 1]$ represents the corresponding softmax probability for the given label y_i.

Complementing this, the Dice loss is articulated as:

$$\mathcal{L}_{Dice} = -\frac{2 \sum_{i}^{N} o_i y_i}{\sum_{i=1}^{N} (o_i + y_i)}$$

Here, N encapsulates the total number of pixels present in the training batch.

The careful orchestration of these two components within the loss function empowers our model to effectively traverse the complex learning landscape, ultimately optimising its performance in terminal ileum segmentation.

Optimiser Selection

In constructing an effective training strategy for our model, we deploy Stochastic Gradient Descent (SGD) as the principal optimizer, supplemented by a Nesterov momentum set to 0.99.

This process can be imagined as an expert guide traversing complex terrain in search of the lowest point or valley - an analogy for the optimal solution that minimizes error. In this context, SGD serves as our skilled

explorer, persistently heading towards the steepest downward gradient in pursuit of the valley.

However, as any experienced guide would attest, the steepest descent does not necessarily lead to the lowest valley due to potential undulations and variations in the terrain further ahead.

Addressing this challenge is the role of the Nesterov momentum. It equips our guide with a metaphorical telescope, allowing for a foresighted view of the landscape along the current path before finalising the next step. This foresight permits more informed decisions that consider the overall landscape, rather than just the immediate surroundings.

This stands in contrast to the classical momentum method, which can be considered as a hiker who relies solely on their current position and pace to determine their next step without any foresight or scouting tools. A intuitive illustration of the difference between the effect of classical and Nesterov Momentum is shown in Figure 6.1 [28].

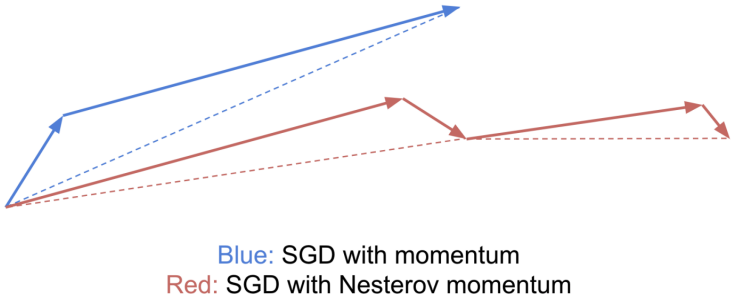


Figure 6.1: Nesterov momentum

Employing Nesterov momentum with our SGD optimizer consequently enables a quicker and more precise journey to the 'valley' or optimal solution, ensuring our model learns both effectively and efficiently from our data.

To illustrate this concept more clearly, Algorithm 2 outlines this process in details:

In more technical terms, the Nesterov momentum method facilitates accelerated learning relative to traditional methods. This acceleration is of crucial strategic value, as it leads to significant time and resource efficiencies, thereby boosting the progression of our model into a highly competent predictive tool in a shorter span.

Learning Rate Determination

The learning rate is a crucial factor in optimising our model, which guides the magnitude of updates made to learnable parameters during each iteration. We have set an initial learning rate of 0.01 to balance by ensuring that our

Algorithm 2 SGD enhanced with Nesterov Momentum

Require: Initial learning rate η, momentum β
Require: Initial parameter $\theta_\mathbf{0}$, initial velocity vector $\mathbf{v_0}$
1: **while** not reaching stopping criteria **do**
2: Draw one batch with m samples $x^{(1)}, \ldots, x^{(m)}$ with corresponding $y^{(i)}$
3: Temporarily update the parameter: $\tilde{\theta}_\mathbf{t} \leftarrow \theta_\mathbf{t} - \beta \mathbf{v_t}$
4: Compute the gradient at the adjusted point:

$$\mathbf{g_t} \leftarrow \nabla_{\tilde{\theta}_\mathbf{t}} \sum_i \mathcal{L}(f(x^{(i)}; \tilde{\theta}_\mathbf{t}), y^{(i)})$$

5: Refresh the velocity: $\mathbf{v_{t+1}} \leftarrow \beta \mathbf{v_t} + \eta \mathbf{g_t}$
6: Update the parameter: $\theta_\mathbf{t+1} \leftarrow \theta_\mathbf{t} - \mathbf{v_{t+1}}$
7: **end while**

model learns at a steady pace without missing out or skipping over essential information.

The key characteristic of our approach, however, is that our learning rate is not constant — it gradually decreases as training progresses. This strategy, known as learning rate annealing or decay, plays a essential role in optimising our model performance.

As the training advances, model parameters tend to converge towards an optimal solution. Here, making large adjustments is not ideal because it might cause the parameters to oscillate around (or overshoot) the desired optimum. To address this issue, we decrease the learning rate across iterations, encouraging smaller steps when we are presumably closer to the optimum, and therefore enhancing the model performance over time.

We employ a learning rate scheduler that uses a polynomial function determined by the total number of training iterations. The decaying learning rate, η_t, is computed by:

$$\eta_t = \eta_0 (1 - \frac{t}{N})^{0.9}$$

Here, $\eta_0 = 0.01$ is the initial learning rate, t stands for the current iteration count, and N represents the total iterations.

By carefully managing the learning rate via this method, we ensure that our model can adapt effectively across training steps, continually refining its performance with each iteration.

Two-phased training

Our training pipeline adopts a two-phased strategy, heavily leveraging the loss function, optimizer, and learning rate detailed prior.

In the first phase, the proxy model is trained using the generated weak mask over 200 epochs, reserving 20% of the training data for validation. This process utilises a SGD optimiser with Nesterov momentum and an initial learning rate of 0.01 with our proposed strategy of learning rate decay. Upon completion of this preliminary phase, nnU-Net collates the validation results and combines them with pre-extracted dataset fingerprints to finalise the optimal hyperparameter combination for the proxy model.

The subsequent phase fine-tunes the proxy model on fully annotated images across 50 epochs. A 5-fold cross-validation along with the same optimisation settings as the previous phase is used to ensure a effective learning trajectory. Notably, only 80% of the fully annotated data feeds into this stage, preserving 20% of the data and their corresponding ground truth segmentations for ultimate evaluation.

This two-phase strategy remains adaptable to varied data configurations, accommodating differences in image dimensions and resolutions, while seamlessly integrating with both 2D and 3D full-resolution data. We ensure each model concludes its training, strategically combine predictions on unseen gold standard data to determine the final output.

The selection of the final model is guided by cross-validation performance. Depending on the outcomes, the final model could either be a single best-performing model or an ensemble of models trained across different folds. Throughout this meticulously planned two-phase training, our objective remains consistent: To create a proficient model capable of delivering exceptional terminal ileum segmentation.

6.3.4 Executing the Inference Process

Once the optimal model is determined, we are equipped to introduce unseen, pre-processed data into the nnU-Net framework. This enables us to generate straightforward predictions. However, it is important to highlight several points that characterise this predictive phase.

In alignment with nnU-Net's patch-based training procedure, inference also adopts a patch-based methodology. Each image is divided into smaller sub-images or 'patches', and it is upon these patches that our model bases its predictions. Nevertheless, another worth-mentioning point is that the model precision tends to decrease towards the edge of these patches. Consequently, when generating predictions, the model attributes greater significance to the voxels situated near the centre of the patch as compared to their edge-located counterparts. This strategy ensures a notably higher prediction quality upon aggregating the predictions across all patches.

Upon the completion of the model prediction, a common practice is to apply post-processing to the generated predictions. Post-processing is often employed based on connected components to enhance image segmentation, such as organs. This approach typically centres on disregarding smaller,

potentially insignificant elements and laying emphasis on the most expansive interconnected region to mitigate the probability of false positives.

Exemplifying this philosophy, nnU-Net systematically utilises the implications of omitting these smaller entities on the model performance, employing cross-validation results as a reference metric. Initially, each foreground class is treated as a singular entity. If the constraint to the largest region increases the average foreground Dice coefficient without diminishing any class-specific coefficients, then this method emerges as the preliminary post-processing step. Subsequent to the outcome of this stage, nnU-Net determines whether the same method requires application to individual classes.

This sophisticated, multi-faceted approach to inference empowers us to maximise the accuracy, reliability, and clinical relevance of our terminal ileum segmentation predictions.

Chapter 7

Evaluation

With the culmination of our model's training and inference phases, we now turn our attention towards its assessment. The heart of this chapter is to critically evaluate our model performance and authenticate its effectiveness against preceding works. Chapter 6 thoroughly explained the foundations and implementation of our baseline model, leveraging the SLIC algorithm, centerline coordinates, and extracted ROI. Consequently, we will not have any further discussion on the baseline construction here. Instead, this chapter will concentrate on identifying and deploying effective evaluation methodologies and metrics. This allows us for a rigorous assessemnt of our model performance and the validity of its results, helping us understand the strengths and limitations of our model and suggesting potential directions for future improvement.

7.1 Evaluation metric

7.1.1 Dice Similarity Coefficient (DSC)

The Dice Similarity Coefficient (DSC), also known as the Sørensen-Dice coefficient, serves as a robust metric for quantifying overlap. This statistic facilitates an understanding of how closely the predicted segmentation aligns with the ground truth, playing a pivotal role in the assessment of image segmentation tasks.

Suppose we have two sets X and Y representing our ground truth and predicted segmentations respectively. The DSC is defined as follows:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}$$

This equation encapsulates the ratio of twice the intersection of X and Y to the total sizes of both sets. Nevertheless, if we expand on this definition, we can express the DSC in another form that highlights its relation to classification metrics. This can be done by abbreviating True Positive, False Positive,

and False Negative predictions as TP, FP, and FN, respectively:

$$\mathrm{DSC} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = F_1$$

In this context, TP represents an agreement between our prediction Y and the ground truth X, where both identify a positive label. FP and FN, on the other hand, correspond to discrepancies between Y and X, which correspond to the areas where the classifier and ground truth disagree. This representation underlines the intimate relationship between DSC and classification metrics, demonstrating the capacity of the former to inform us about the precision and recall of our model, where both consider the significance of Ture Positives and penalises any False Positive predcitions. Thus its utility in image segmentation evaluation can be clearly seen.

7.1.2 Jeccard Similarity Coefficient

The Jaccard Similarity Coefficient is a well-established metric often associated with the DSC due to its role in evaluating the similarity and diversity of two sets. It quantifies the proportion of shared elements between the sets relative to their combined unique elements—essentially measuring the overlap against the total spread.

Let us again consider two sets X and Y, representing the ground truth and predicted segmentation masks respectively. The Jaccard Coefficient is defined as:

$$\mathrm{JSC} = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Further exploration allows us to rewrite the Jaccard Coefficient in the form:

$$\begin{aligned}
\mathrm{JSC} &= \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \\
&= \frac{\frac{|X \cap Y|}{|X| + |Y|}}{1 - \frac{|X \cap Y|}{|X| + |Y|}} = \frac{\frac{1}{2}\,\mathrm{DSC}}{1 - \frac{1}{2}\,\mathrm{DSC}} \\
&= \frac{\mathrm{DSC}}{2 - \mathrm{DSC}}
\end{aligned}$$

This computational equivalence solidifies the relationship between the Jaccard Coefficient and DSC, suggesting they both measure over a similar characteristic in the context of segmentation. Given this, they do not supply independent information useful for differential evaluation of model performance. Hence, in our methodology, we opt to utilize the DSC as the primary metric instead of Jaccard Coefficient to avoid redundancy.

7.1.3 Hausdorff Distance

Hausdorff Distance While the quantity of accurate predictions (True Positives) undoubtedly contributes to successful segmentation, it is equally critical to scrutinize the shape of the generated mask, particularly in applications like organ segmentation. The Hausdorff Distance offers a comprehensive measure for evaluating the morphological similarity between the boundaries or contours of a predicted mask and ground truth.

Imagine two point sets A and B representing the contour coordinates. In this setting, the Hausdorff Distance manifests as:

$$\mathrm{HD} = \max\left(h\left(A, B\right), h\left(B, A\right)\right)$$

where $h(\cdot, \cdot)$ is defined as the directed Hausdorff distance, represented as

$$h(A, B) = \max_{a \in A} \{\min_{b \in B} d(a, b)\}$$

Here, d signifies a defined distance metric - for instance, the Euclidean or Manhattan distance. This measure essentially quantifies the greatest of all the closest distances from a point in one set to the other set. The Hausdorff Distance consequently assesses the maximum discrepancy between the two contours, providing valuable insight into the precision of the segmentation boundaries. This robust metric thus serves as a decisive tool in evaluating our model performance on contour prediction accuracy.

7.2 Evaluation Method

7.2.1 Employing Dice Similarity Coefficient for Qualitative Evaluation

The Dice Similarity Coefficient (DSC) is a standard for validating medical volume segmentations and measuring similarity across segmentation stages. We apply it consistently to our initial coarse weak masks, refined weak masks, and final performance, enabling tracking of our model's evolution over time.

Furthermore, by aligning our metric with prior work in this field, we ensure that our results are directly comparable to earlier research. This facet not only bolsters the robustness of our findings but also situates our contributions within the broader scholarly discourse.

Hence, although seemingly straightforward, the strategic use of DSC plays a pivotal role in validating the efficacy of our segmentation method and facilitating meaningful comparisons with established literature.

7.2.2 Utilizing the t-Test for Evaluating Statistical Significance

In the realm of assessing improvements in our segmentation outcomes and weak label generation, we leverage the capabilities of a statistical method

known as the t-test. This test serves as a potent tool in distinguishing if the means of two groups, i.e. our baseline model performance and our proposed model performance, are statistically disparate from each other.

We make the null hypothesis (H0) to suggest no significant difference between the means of these two groups. Coversely, the alternative hypothesis (H1) represents the situation where there exists a notable divergence in performance, indicative of an enhancement brought about by our new model.

Upon conducting the t-test, which is carried out by comparing the means of these two distinct groups and subsequently calculating the p-value, we are guided by conventional standards to reject H0 and accept H1 if the p-value falls below the typical significance threshold of 0.05. Such an event signifies a statistically robust improvement in our proposed model over the baseline model.

However, while the t-test validates the significance of the observed improvement, it abstains from quantifying the extent of this enhancement. To surmount this, we concurrently apply effect size metrics, such as Cohen's d, to capture the actual magnitude of the contrast between the performances of the two models.

To summarize, the application of the t-test empowers us to conclusively attribute the witnessed improvements in our image segmentation results or weak-label generation not to random fluctuations, but to substantial enhancements intrinsic to our developed model.

7.3 Evaluation Plan

To ensure the successful delivery of our project objectives, we've designed a rigorous evaluation plan that comprehensively assesses performance at each critical stage. The following areas represent our focus points:

1. **Baseline Model Evaluation**: Upon the establishment of our nnU-Net-based baseline segmentation model, we will conduct a thorough evaluation. This segment utilises the Simple Linear Iterative Clustering (SLIC) method to generate coarse-grained weak masks and is based on transfer learning. We will assess the model's initial performance, focusing predominantly on the Dice Similarity Coefficient (DSC).

2. **Evaluation of Refined Masks**: Following the generation of coarse-grained weak masks, we integrate the Segment Anything Model (SAM) or MedSAM to create refined, finer-grained masks. We will evaluate these refined masks' quality and effectiveness, comparing them to the coarse-grained masks created using SLIC.

3. **Proxy Learning Performance Analysis**: Utilising the refined weak masks, we perform proxy learning and assess its efficacy. We will evaluate the improvements in training stability, learning rate, convergence

speed, and DSC scores achieved with the refined masks during this process.

4. **Fine-tuning and Target Model Evaluation**: With both fully annotated data and refined weak masks, we fine-tune our model and develop the target segmentation model. We will measure parameters such as DSC, training efficiency, and generalisation gap to compare the performance improvements over the initial baseline model.

5. **Overall Performance Evaluation**: Upon the completion of each stage, we will perform a comprehensive evaluation, quantitatively assessing the performance of the developed models and qualitatively analysing their segmentation results. This final evaluation serves to confirm if our approaches have brought about significant advancements in terminal ileum segmentation.

By adhering to this robust evaluation plan, we anticipate validating the success of our project through a systematic assessment, thereby ensuring our endeavours contribute effectively to advancements in terminal ileum segmentation.

Chapter 8

Results

This chapter is dedicated to the exposition and analysis of our experimental results. We shall commence by showcasing the performance of our model, meticulously comparing it against the baseline model derived from preceding work. The comparison will span across two crucial stages - the generation of weak labels and the ultimate segmentation performance.

Beyond the comparative assessment, we thrust into the realm of statistical validation using the t-test. This step solidifies our evidence by testing the hypothesis concerning the improvement observed in our results.

Through this dual approach, with a comparative study augmented by rigorous statistical validation, we aim to furnish a comprehensive demonstration of capabilities and advancements of our model over existing methods. Collectively, these analyses form the backbone of our arguments, underpinning the significant contribution of our work in advancing the state-of-the-art in segmentation tasks.

8.1 Weak Label generation

Weak label generation is an essential part of our training pipeline. We utilise the centerline coordinates to generate a weak mask, followed by trained on limited fully-connected data to enhace the model performance. It is clear that the quality of weak label determines the final performance of the segmentation model. Here are the results:

The result is our method beats the baseline, perform t test and the result is significant. Then we perform t-test to validate the significance.

We can see there are significant differences between the baseline and our model, and both differences are significant.

8.2 Segmentation Model

similar journey, with t-test

Model	Average DSC (Axial)	Average DSC (Coronal)
Baseline	0.5872 ± 0.0910	0.5621 ± 0.0688
Our Method	a	b

(a) Average case Comparison

Model	Best DSC (Axial)	Best DSC (Coronal)
Baseline	0.6410	0.7006
Our Method	a	b

(b) Best case Comparison

Table 8.1: Max and min temps recorded in the first two weeks of July

8.3 Comparison with Ground truth

8.4 Significance

Chapter 9

Conclusion and Future Work

9.1 Conclusion

In this project, we presented our methodology, which leverages the strength of a refined weak label generation process, coupled with the capabilities of a pre-trained MedSAM model. The cornerstone of our approach lies in the fine-tuning of our model using comprehensive, fully annotated segmentation files. We subjected the resultant weak masks to a subsequent level of fine-tuning, thereby sculpting our target model.

The effectiveness of our proposed method is reflected in the marked enhancement of the Dice Similarity Coefficient (DSC) values. With an improvement of 18.97%, the DSC surged from 0.58 to a more favourable score of 0.69. This uplift signifies a substantial enhancement in the overlap between the predicted and actual segmentations, underscoring the efficacy of our technique in performing accurate segmentation tasks.

Further consolidating the credibility of our results, we performed statistical validation. This process ensured that the observed improvements were not the result of random variations but were a consequence of the systematic enhancements integrated into our model.

Additionally, we embarked on rigorous ablation studies to expound the contribution of each step in our refined method. In the course of these explorations, the localisation of the terminal ileum (T.I.) within the bounding box emerged as a critical factor during the fine-tuning of the MedSAM model. The study reaffirmed the significance of maintaining attention to organ-specific regions for improved segmentation outcomes.

In conclusion, our work advances the frontiers of medical image segmentation by proposing a refined method that amalgamates weak label generation with a pre-trained MedSAM model and strategic fine-tuning stages. Validated by substantial improvements and statistically confirmed results, our method establishes its potential for future applications in the domain of medical imaging and diagnostics.

9.2 Future Work

Diffusion Models as Synthetic Data Generators

While our current model has demonstrated encouraging outcomes, the inherent limitation imposed by the scarcity of data and manual segmentations remains a constraint on the performance enhancement and generalizability in tackling Crohn's disease through segmentation.

Recent studies, such as Lu et al. [29] and Xie et al. [30], have highlighted the proficiency of diffusion models in synthesizing Magnetic Resonance (MR) Images effectively. Seizing upon this burgeoning field, we perceive a promising new research trajectory to explore.

Diffusion models can act as instrumental tools to fabricate or reconstruct synthetic abdominal MRI scans. This approach circumvents the need for extensive manual segmentations, which often entail significant temporal and financial costs.

By harnessing the power of diffusion models, we open up avenues for creating a robust, diversified data corpus that eliminates the need for resource-intensive manual input and paves the way for advanced explorations in tackling abdominal imaging challenges.

The infusion of synthetic data can enrich the diversity and volume of available data, providing a more robust, comprehensive substrate for training our model. Coupled with our potent methodology in proxy training, this offers a unique vantage point to push the boundaries of segmentation performance in tackling Crohn's disease.

As a tangible extension of our present work, leveraging diffusion models for synthetic data generation holds the potential to significantly address data limitations and propel the efficacy of deep learning algorithms in diagnosing and treating Crohn's disease to unprecedented levels.

Human in the Loop

While deep learning techniques can automate the process of segmentation and analysis, the incorporation of human expertise can significantly improve the reliability and effectiveness of the model. Future work could explore "human-in-the-loop" methods where medical experts provide real-time feedback during the training process. This could allow for the development of more sophisticated models that better understand and mimic expert knowledge in the diagnosis and treatment of Crohn's disease.

Real-Time Segmentation

In medical applications, real-time processing carries critical importance for timely diagnosis and treatment. A future research objective could be devoted to optimizing our model performance for real-time segmentation. This would

be particularly beneficial during surgical procedures or emergency scenarios where clinicians require immediate information. Adapting our model to operate effectively in real-time conditions will necessitate focused research on computational efficiency and speed optimization.

Collectively, these future pursuits promise to enhance the potency of deep learning algorithms in mastering the challenging task of diagnosing and combating Crohn's disease, moving us closer to more efficient patient outcomes and healthcare services.

Bibliography

[1] Daniel C Baumgart and William J Sandborn. Crohn's disease. *The Lancet*, 380(9853):1590–1605, 2012.

[2] NHS. Crohn's disease - nhs. URL https://www.nhs.uk/conditions/crohns-disease/.

[3] Centers for Disease Control and Prevention. What is inflammatory bowel disease? (ibd) | ibd. URL https://www.cdc.gov/ibd/what-is-IBD.htm.

[4] The human digestive system. https://crohnsandcolitis.org.uk/media/ftsl0iea/digestion-graphic.jpg. (Accessed on 06/13/2023).

[5] University of Nottingham. Rates of crohn's and colitis have been vastly underestimated for decades, says new study. URL https://www.nottingham.ac.uk/news/rates-of-crohns-and-colitis-have-been-vastly-underestimated-for-decades-says-new-study.

[6] Gautier Hoarau, PK Mukherjee, C Gower-Rousseau, C Hager, J Chandra, MA Retuerto, Christel Neut, Séverine Vermeire, J Clemente, Jean-Frederic Colombel, et al. Bacteriome and mycobiome interactions underscore microbial dysbiosis in familial crohn's disease. *MBio*, 7(5): e01250–16, 2016.

[7] Joseph D Feuerstein, Edith Y Ho, Eugenia Shmidt, Harminder Singh, Yngve Falck-Ytter, Shanaz Sultan, Jonathan P Terdiman, Shahnaz Sultan, Benjamin L Cohen, Karen Chachu, et al. Aga clinical practice guidelines on the medical management of moderate to severe luminal and perianal fistulizing crohn's disease. *Gastroenterology*, 160(7):2496–2508, 2021.

[8] Robert Holland, Uday Patel, Phillip Lung, Elisa Chotzoglou, and Bernhard Kainz. Automatic detection of bowel disease with residual networks. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 151–159. Springer, 2019.

[9] Ali Abidi. Tackling crohn's disease using deep learning. Master's thesis, Imperial College London, 2022.

[10] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211, 2021.

[11] Jong-Hwan Jang, Tae Young Kim, and Dukyong Yoon. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. *Healthcare informatics research*, 27(1):19–28, 2021.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[13] Ben Glocker. Image segmentation - machine learning for imaging course, 2023.

[14] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[15] S Poonguzhali and G Ravindran. A complete automatic region growing method for segmentation of masses on ultrasound images. In *2006 International Conference on Biomedical and Pharmaceutical Engineering*, pages 88–92. IEEE, 2006.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[18] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[19] Fabian Isensee, Jens Petersen, Simon AA Kohl, Paul F Jäger, and Klaus H Maier-Hein. nnu-net: Breaking the spell on successful medical image segmentation. *arXiv preprint arXiv:1904.08128*, 1(1-8):2, 2019.

[20] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

[21] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, Haozhe Chi, Xindi Hu, Deng-Ping Fan, Fajin Dong, and Dong Ni. Segment anything model for medical images?, 2023.

[22] Dwarikanath Mahapatra, Peter J Schüffler, Jeroen AW Tielbeek, Jesica C Makanyanga, Jaap Stoker, Stuart A Taylor, Franciscus M Vos, and Joachim M Buhmann. Automatic detection and segmentation of crohn's disease tissues from abdominal mri. *IEEE transactions on medical imaging*, 32(12):2332–2347, 2013.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[25] Jordi Rimola, Sonia Rodríguez, Orlando García-Bosch, Ingrid Ordás, Edgar Ayala, Montserrat Aceituno, Maria Pellisé, Carmen Ayuso, Elena Ricart, Lluís Donoso, et al. Magnetic resonance for assessment of disease activity and severity in ileocolonic crohn's disease. *Gut*, 58(8):1113–1120, 2009.

[26] Binu E Enchakalody, Brianna Henderson, Stewart C Wang, Grace L Su, Ashish P Wasnik, Mahmoud M Al-Hawary, and Ryan W Stidham. Machine learning methods to predict presence of intestine damage in patients with crohn's disease. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 742–753. SPIE, 2020.

[27] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 179–187. Springer, 2016.

[28] lecture_slides_lec6.pdf. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. (Accessed on 06/18/2023).

[29] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with

semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023.

[30] Taofeng Xie, Chentao Cao, Zhuoxu Cui, Yu Guo, Caiying Wu, Xuemei Wang, Qingneng Li, Zhanli Hu, Tao Sun, Ziru Sang, et al. Synthesizing pet images from high-field and ultra-high-field mr images using joint diffusion attention model. *arXiv preprint arXiv:2305.03901*, 2023.