AN INVESTIGATION OF AUTOMATED MODELS FOR TUMOR SEGMENTATION

AN INVESTIGATION OF ADVANCED DEEP LEARNING-BASED AUTOMATED MODELS FOR TUMOR SEGMENTATION IN WHOLE-BODY PET/CT IMAGES

By MAHAN POUROMIDI, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Applied Sciences

McMaster University
MASTER OF APPLIED SCIENCES (2024)
Hamilton, Ontario, Canada (School of Biomedical Engineering)

TITLE:	An investigation of advanced deep learning-based automated models for tumor segmentation in whole-body PET/CT images
AUTHOR:	Mahan Pouromidi B.Sc. (Electrical Engineering), Amirkabir University of Technology, Tehran, Iran
SUPERVISOR:	Dr. Ashirbani Saha
NUMBER OF PAGES:	xxii, 135

Abstract

In this work, we focus on the segmentation of tumors on PET/CT [Positron Emission Tomography used with Computed Tomography], which is crucial in routine clinical oncology. Based on the advances in recent deep learning-based methodologies, we studied the relative performances of three different frameworks: (a) nnU-Net [Convolutional Neural Network (CNN)-based], (b) nnU-Net with prompting a large Vision-Transformer (ViT) model called Segment Anything Model (SAM) (Hybrid), and (c) Swin-Unet (U-Net-like pure transformer) in a publicly available dataset of PET/CT images including normal patients and patients with lung cancer, lymphoma, and melanoma. Our study includes a holistic performance analysis for three cancer types and normal cases, which is typically avoided in the literature. The image volumes with cancer typically include more than one lesion (primary tumor and potential metastases). Therefore, we conducted two types of analyses. Our first analysis is conducted at an image volume level, considering all lesions together as foreground, and the rest as background. For the second analysis, we executed connected-component labelling to algorithmically label different parts of the tumor and assessed at lesion component level. At image volume level, nnU-Net performed best for lung cancer (Dice score: 73.25%) compared to melanoma (63%) and lymphoma (72.6%) among the three methods. The median largest lesion component-wise Dice score for nnU-Net, SAM with

nnU-Net prompts, and Swin-Unet on three cancer types combined are 85%, 67%, and 72%, respectively. Both nnU-Net and SAM with nnU-Net approaches missed 2, 4, and 4 image volumes of lung cancer, lymphoma, and melanoma patients, resp., whereas Swin-Unet did not miss a single volume. Out of 513 normal volumes, 201 were successfully identified by nnU-Net and SAM, whereas Swin-Unet only identified 7 of them. In conclusion, the performance of models varied across the cancer types. nnU-Net proved to be the most reliable and precise algorithm evaluated in this study by showing the best performance for identifying normal patients and in delineating the largest lesions.

To all the love that I received during this journey

Acknowledgements

I want to express my profound gratitude to my supervisor, Dr. Ashirbani Saha, from whom I have learned immensely. Her guidance extended beyond academic knowledge to invaluable life lessons. Her perseverance and dedication have been a great source of inspiration throughout this journey.

Furthermore, I extend my thanks to CREATE (CentRE for dAta science and digiTal hEalth), Hamilton Health Sciences, and PHRI (Population Health Research Institute) for providing the facilities and computing resources essential for our high-performance computing tasks. Special thanks to Walter Nelson for his insights on troubleshooting our models, which greatly enhanced this project.

I also thank the committee members, Dr. Thomas Doyle and Dr. Katherine Zukotynski, for their guidance and insights that improved this research work.

I sincerely acknowledge the efforts of Maya Sabados, the program coordinator of the School of Biomedical Engineering at McMaster University.

Last but certainly not least, I thank my family members for their unwavering emotional support overseas, which helped me through the challenging times during my master's studies.

Table of Contents

Abstract **iii**

Acknowledgements **vi**

Definitions and Abbreviations **xix**

Declaration of Academic Achievement **xxiii**

1 Introduction **1**

1.1 General overview of image segmentation 3

1.2 Medical image segmentation 6

1.3 Importance of PET/CT image segmentation 6

1.4 Deep learning and its advances for imaging analysis 11

1.5 Problem statement 12

1.6 Organization of the thesis 12

2 Literature Review and Related Concepts **14**

2.1 Searching methodology 14

2.2 Image processing and computer vision-based methods 15

2.3 Traditional machine learning 17

2.4 Deep learning 18
2.5 Drawbacks of current studies in PET/CT segmentation 31

3 Materials and Methods **33**
3.1 Dataset 33
3.2 Data pre-processing 35
3.3 nnU-Net 37
3.4 Segment Anything Model (SAM) with nnU-Net prompts 44
3.5 Swin-Unet 50
3.6 Evaluation metrics 54
3.7 Analysis paradigm 55

4 Results **58**
4.1 Volume-wise analysis 58
4.2 Lesion component-wise analysis 68

5 Discussion **72**

6 Limitations **75**

7 Conclusion and Future Works **78**

A **80**
A.1 nnU-Net model evaluation on partial dataset 80
A.2 Performance metrics comparison for SAM and Swin-Unet with respect to cancer type in whole-body volumes 85
A.3 Further insight on the dataset and nnU-Net's performance 90

B **115**

B.1 Code for connected component analysis 115

B.2 Code for random dilation 115

B.3 Code for CT image pre-processing 116

B.4 Code for SUV contrast stretching 117

List of Figures

1.1 Differences between image classification, object localization, semantic segmentation, and instance segmentation. Adopted from[1] 5
2.1 U-Net architecture. Adopted from[2] 20
2.2 Transformer architecture. Adopted from[3] 29
2.3 Vision transformer architecture for classification task. Adopted from[4] 30
3.1 A slice from a segmentation volume map, containing five 2D connected components 36
3.2 nnU-Net pipeline. Adopted from[5] 37
3.3 nnU-Net development procedure. Adopted from[5] 37
3.4 Comparison of dilated segmentation map (left) versus the original segmentation map (right). Adopted from[6] 47
3.5 Box prompting in SAM. Adopted from[7] 49
3.6 Point prompting in SAM. Adopted from[7] 49
3.7 SAM architecture. Adopted from[8] 49
3.8 Swin-Unet architecture. Adopted from[9] 53
4.1 Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on lung cancer patients using five performance measures. 'Dice' denotes the Dice score. 59

4.2 Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on lymphoma patients using five performance measures. 'Dice' denotes the Dice score. 59
4.3 Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on melanoma patients using five performance measures. 'Dice' denotes the Dice score. 60
4.4 Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on three cancers combined using five performance measures. 'Dice' denotes the Dice score 60
4.5 Comparison of cancer evaluation metrics for the nnU-Net method. 'Dice' denotes the Dice score, and stars indicate the best scores. . . . 61
4.6 Dice score versus overall tumor volume 63
4.7 Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all lung cancer patients. This slice has the largest tumor volume of all slices for this patient. The SAM experiment uses 3 PET channels and no dilation . 65
4.8 Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all lymphoma patients. This slice has the largest tumor volume of all slices for this patient. The SAM experiment uses 3 PET channels and no dilation . 66
4.9 Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all melanoma patients. This slice has the largest tumor volume of all slices for this patient. The SAM experiment uses 3 PET channels and no dilation . 67

4.10 Spearman correlation between the number of connected components in the ground truth and the number of connected components in the prediction map 68
A.1 Modality and loss function performance comparison for the partial dataset using nnU-Net 84
A.2 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 3 PET channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score. 85
A.3 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 3 PET channels and dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score. 86
A.4 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 2 PET, 1 CT channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score. 86
A.5 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 2 PET, 1 CT channels and dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score. 87
A.6 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 1 PET, 2 CT channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score. 87
A.7 Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 1 PET, 2 CT channels and dilation" method. 'Dice' denotes the Dice score, with and stars indicate the highest score. . . . 88

A.8 Comparison of cancer evaluation metrics for the Swin-Unet method. 'Dice' denotes the Dice score, with and stars indicate the highest score. 89
A.9 Distribution of SUVs corresponding to PET-positive malignancy (presence of cancer) across three cancer types in ground truth. 90
A.10 Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in lung cancer patients for nnU-Net algorithm. 91
A.11 Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in lymphoma patients for nnU-Net algorithm. 92
A.12 Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in melanoma patients for nnU-Net algorithm. 93
A.13 Histogram of volumes of all the lesions (ground truth) in lung cancer patients for nnU-Net algorithm. 94
A.14 Histogram of volumes of all the lesions (ground truth) in lymphoma patients for nnU-Net algorithm. 95
A.15 Histogram of volumes of all the lesions (ground truth) in melanoma patients for nnU-Net algorithm. 96
A.16 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lung cancer patients. 97

A.17 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lung cancer patients with limited x-axis for a better view on the low volume lesions. 98
A.18 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lymphoma patients. 99
A.19 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lymphoma patients with limited x-axis for a better view on the low volume lesions. 100
A.20 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in melanoma patients. 101
A.21 Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in melanoma patients with limited x-axis for a better view on the low volume lesions. 102

A.22 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lung cancer patients. 103
A.23 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lung cancer patients with limited x-axis for a better view on the low volume lesions. 104
A.24 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lymphoma patients. 105
A.25 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lymphoma patients with limited x-axis for a better view on the low volume lesions. 106
A.26 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in melanoma patients. 107

A.27 Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in melanoma patients with limited x-axis for a better view on the low volume lesions. 108
A.28 Percentage of each lesion that has been correctly identified with respect to their volumes in lung cancer patients for nnU-Net algorithm. . . . 109
A.29 Percentage of each lesion that has been correctly identified with respect to their volumes in lung cancer patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm. 110
A.30 Percentage of each lesion that has been correctly identified with respect to their volumes in lymphoma patients for nnU-Net algorithm. 111
A.31 Percentage of each lesion that has been correctly identified with respect to their volumes in lymphoma patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm. 112
A.32 Percentage of each lesion that has been correctly identified with respect to their volumes in melanoma patients for nnU-Net algorithm. 113
A.33 Percentage of each lesion that has been correctly identified with respect to their volumes in melanoma patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm. 114

List of Tables

3.1 Dataset summary 35

4.1 Number of missed image volumes by cancer type. 'SAM' refers to all the six "SAM with nnU-Net prompts" approaches. 62

4.2 Number of normal image volumes successfully flagged as normal . . . 62

4.3 Median average lesion component-wise Dice score (Range). Bold numbers indicate the highest values obtained for each cancer type and combined. D: dilation, ND: no dilation 70

4.4 Median highest lesion component-wise Dice score (Range). Bold numbers indicate the highest values obtained for each cancer type and combined. D: dilation, ND: no dilation 70

4.5 Median volume-wise Dice score (Range). Bold numbers indicate the highest values obtained for each cancer type and combined. D: dilation, ND: no dilation 70

4.6 Percentage of image volumes for which the highest lesion component-wise Dice score was obtained from the largest lesion component in the ground truth. Bold numbers indicate the highest values obtained for each cancer type and combined. D: dilation, ND: no dilation 71

4.7 Number of lesion-component pairs (lesion-component from ground truth versus that from predicted segmentation map) for which the calculated Dice scores are not zero. D: dilation, ND: no dilation 71
A.1 Train/test split in partial dataset for nnU-Net 81
A.2 Normal cases successfully identified using PET only (out of 25) 81
A.3 Malignant cases identified as normal using PET only (out of 111) . . 81
A.4 Performance evaluation metrics for partial dataset trained on PET only, evaluated on the entire dataset 81
A.5 Performance evaluation metrics for partial dataset trained on PET only, evaluated on positive volumes only 82
A.6 Normal cases successfully identified using CT only (out of 25) 82
A.7 Malignant cases identified as normal using CT only (out of 41) 82
A.8 Performance evaluation metrics for partial dataset trained on CT only, evaluated on the entire dataset 82
A.9 Performance evaluation metrics for partial dataset trained for CT only, evaluated on positive volumes only 82
A.10 Normal cases successfully identified using PET/CT (out of 25) 83
A.11 Malignant cases identified as normal using PET/CT (out of 41) . . . 83
A.12 Performance evaluation metrics for partial dataset trained on PET/CT, evaluated on the entire dataset 83
A.13 Performance evaluation metrics for partial dataset trained on PET/CT, evaluated on positive volumes only 83

Definitions and Abbreviations

Definitions

Image processing

Refers to the general set of computational methods for analyzing and manipulating images to extract patterns and meaningful data, which would take longer time or much effort if done manually.

Medical Image Segmentation

The process of manipulating a digital image related to human's body to identify a specific region-of-interest by delineating the detailed boundaries of a desired region. This could be done manually, semi-manually (also called semi-automatically), or automatically.

Deep Learning

An advanced category of machine learning methodologies that employs multi-layered neural networks to iteratively extract increasingly sophisticated features from raw data inputs.

CNN Convolutional Neural Network, a deep learning algorithm engineered to analyze input images, assigning significance to various features or

objects within the image through learnable weights and biases, and effectively differentiating between them. The term 'convolutional' pertains to the process of sliding filters across the image using am athematical operation called 'convolution operation'.

Transformers An architecture for transforming an input sequence to a desired output with the help of two parts (Encoder and Decoder), using a mechanism called "self-attention". It processes all the input elements in parallel irrespective of their position (unlike traditional sequential models) allowing for more efficient results.

Image volume

A 3D image representation of human anatomy consisting of cross-sectional views (2D slices) stacked along a specific direction.

Abbreviations

2D	Two dimensional
3D	Three dimensional
AI	Artificial intelligence
CNN	Convolutional neural network
CT	Computed tomography
FDG	Fluorodeoxyglucose
FFN	Feed-Forward Networks

GAN	Generative adversarial network
GPU	Graphics processing unit
GTV	Gross tumor volume
KV	Kilovolts
LSTM	Long short-term memory
mAs	Milliampere-seconds
MBq	Megabecquerel
MR	Magnetic resonance
MRF	Markov random field
NLP	Natural language processing
NPC	Nasopharyngeal carcinoma
NSCLC	Non-small cell lung cancer
OCS	Optimum contour selection
ONNX	Open neural network exchange
OSEM	Ordered-subset expectation maximization
PET	Positron emission tomography
PET-CT	Positron emission tomography - Computed tomography
ReLU	Rectified linear unit

ROI	Region of interest
SAM	Segment anything model
STS	Soft tissue sarcoma
SUV	Standard uptake value
SVM	Support vector machine
ViT	Vision transformer

Declaration of Academic Achievement

This declaration confirms that the research conducted in this thesis was completed by Mr. Mahan Pouromidi and recognizes the contributions of Dr. Ashirbani Saha. Mahan Pouromidi led the data processing, analysis, experiments, coding, troubleshooting, and thesis drafting. Dr. Ashirbani Saha contributed to the inception of the study, analysis, supervision, and review of the thesis.

Chapter 1

Introduction

Cancer's incidence continues to rise at an alarming rate worldwide[10]. The urgency of innovative solutions in cancer care is highlighted by the alarming prevalence of cancer worldwide and in Canada. Throughout the cancer continuum, biomedical or medical imaging plays an important part[11]. A variety of medical imaging modalities, such as computed tomography, x-ray, magnetic resonance (MR) imaging, etc., are used for cancer screening, diagnosis, monitoring, treatment, follow-up, and drug discovery[12]. The physical properties (e.g., sensitivity, temporal and spatial resolution) of different medical imaging modalities can vary significantly.

Methods for semi-automated and automated analysis of different medical imaging modalities[13][14][15][16] have existed for more than several decades now. These methods are based on image processing/computer vision[17], machine learning[18], and more recently, deep learning[19], and data-driven models in Artificial Intelligence (AI)[20]. Examples of these image analyses include image registration[21], image segmentation[22], image classification for several tasks such as diagnosis[23], disease types[24], and organs[25].

This research work is related to image segmentation. Image segmentation involves detailing the outline of regions of interest (ROIs) in an image manually, semi-automatically, or automatically through computational techniques. Despite decades of seminal work, the task of automated medical image segmentation remains an unsolved problem due to its complexity.

Several factors contribute to this complexity. Firstly, defining the problem is challenging, as ground truth or reference standard ROIs exhibit intra- and inter-reader variability. Secondly, data complexity is high; different medical imaging modalities have unique properties, and even the same modality can vary in appearance and properties based on the image acquisition protocol and equipment. Additionally, medical images vary significantly in dimensionality—ranging from 2D, 3D, to 4D (3D + time)—and resolution. The non-rigid nature and deformation of human body parts also introduce variation.

Nevertheless, automated image segmentation is highly sought after to improve efficiency in various tasks, such as diagnosis through abnormality detection and subsequent outlining, automated measurements in images, treatment planning, and disease management[26][27].

Automated image segmentation in cancer is particularly significant as it can expedite the screening and measurement of lesion properties (including volume, shape, and textural properties or quantitative imaging biomarkers), assist in treatment planning, and enable the classification of lesions (e.g., benign/malignant, histological grades in lung cancer, molecular subtypes in breast cancer). Although image segmentation is highly dependent on the imaging modality, the advent of new deep learning techniques has mitigated this dependence. However, in contrast to non-medical or natural

images, medical images (such as radiological and pathological images) remain more modality-dependent despite the emergence of recent foundational models that support a variety of imaging modalities[28].

The work presented in this master's thesis involves the implementation and detailed evaluation of three recent deep learning-based techniques for PET/CT segmentation using a publicly available dataset from The Cancer Imaging Archive (TCIA)[29]. This dataset includes image volumes (defined as a stack of 2D slices) from patients with various cancer types, with multiple lesions, as well as images from normal patients. The results presented in the thesis describe the quantitative evaluation of three recent competitive deep learning-based techniques both overall and by cancer type. The evaluation is conducted in two ways: (a) using the traditional image volume-wise evaluation method, and (b) through an unconventional lesion-specific analysis. The latter is important for subsequent tasks such as various measurements from lesions and feature extraction, which occur after segmentation but are not strongly emphasized in the literature. The rest of this chapter focuses on providing the background of automated PET/CT segmentation and offers an overview of recent advances in deep learning literature. This sets the stage for the subsequent description of the problem statement of this thesis.

1.1 General overview of image segmentation

Image segmentation involves dividing an image into fragments (ROIs), or segments, and assigning a label to each segment based on the content. This process can sometimes be confused with similar concepts such as image classification or object localization.

The objective of image classification is to determine the presence of specific objects within an image, typically producing probabilities indicating presence of those, or identifying a single main object type (such as cups, bottles, etc.) in the image[1].

In object localization, the goal is to locate all instances of predefined objects within an image and predict bounding boxes and labels for each object[1].

Semantic segmentation, a type of image segmentation, aims to label all the pixels in an image. For example, in Figure 1.1 on the bottom left, the pixels that belong to any type of cube would be assigned an integer label (let's say 1 in this case), and all the pixels representing cups in the image would be assigned another label (let's say 2). The objective illustrated in this figure is to assign labels to specific objects (cubes, bottles, and cups in this example). Anything else would be regarded as the "background" and would receive a label of 0.

Instance segmentation is a sub-division of image segmentation that advances beyond merely labeling every pixel of each object by also differentiating between distinct instances of the same class. For example, all the cubes in the figure below would be labeled separately, even though they all belong to the same class, allowing each cube to be uniquely identified.

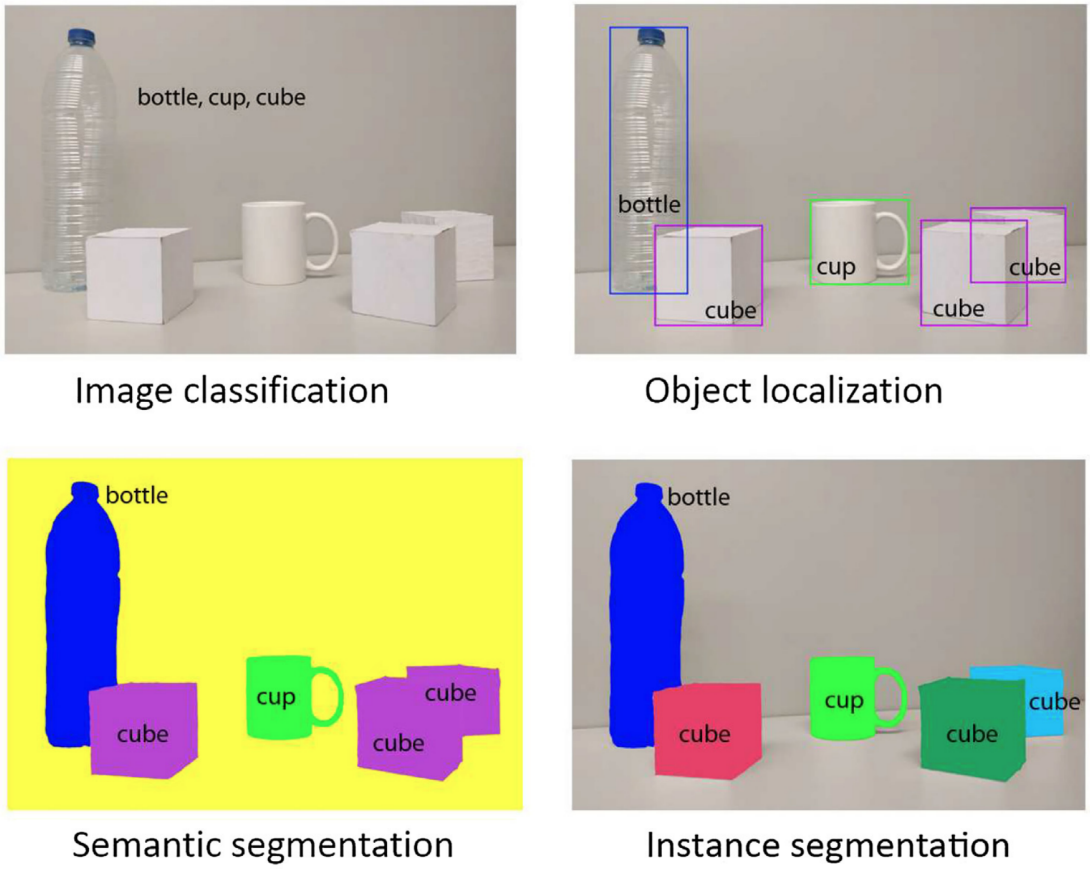

Figure 1.1: Differences between image classification, object localization, semantic segmentation, and instance segmentation. Adopted from[1]

Image segmentation has been utilized in various fields, and one significant application is in healthcare, where medical experts take advantage of the resulting extracted data for better disease diagnosis, treatment planning, and more precise measurements.

1.2 Medical image segmentation

Medical images encompass essential biological data, serving as a crucial foundation for healthcare professionals to formulate informed decisions. Handling such images fundamentally differs from managing conventional images captured by standard cameras, as they entail various modalities with different dynamic ranges. These images carry meaningful biological concepts, so interpretability is of utmost importance. They are also harder to obtain, making it more challenging to gather enough data for automated models, which require a plethora of data to perform well compared to manual methods. There are multiple modalities using which experts and developers utilize to perform segmentation tasks on organ, tumors, or tissues[28].

1.3 Importance of PET/CT image segmentation

PET/CT plays a pivotal role in oncology by providing valuable information for cancer diagnosis, staging, treatment planning, and monitoring response to therapy. The integration of metabolic information from PET with anatomical details from CT enhances the precision and accuracy of cancer assessment. PET/CT is particularly vital in identifying primary tumor locations, detecting metastases, and evaluating treatment efficacy. The ability to visualize metabolic activity aids in distinguishing between benign and malignant lesions, facilitating more informed clinical decisions[30].

We will now discuss the underlying physics of PET/CT and how the images are acquired.

1.3.1 Physics of PET/CT acquisition

PET is a nuclear medicine imaging technique that utilizes radioactive tracers to monitor biochemical processes in the body. The tracers, composed of carrier molecules tightly linked to radioisotopes (unstable isotopes), are introduced to human body via intravenous injection. These carrier molecules selectively interact with specific proteins or sugars, chosen based on medical professionals' diagnostic objectives. The stabilization of radioisotopes can occur through positron emission or electron capture, with elements having lower atomic weight typically favoring positron emission for stabilization.

Following the administration of the radiotracer, the emitted positron travels a short range in the body, leading to two possible scenarios: combining with a free electron in the electron cloud, producing a pair of 511 keV photons (also referred to as "annihilation photons"), or forming a positronium, which decays into a pair of 511 keV photons. Detectors around the patient's body capture the annihilation photons, illuminating corresponding areas in the final image.

Although the acquired image provides valuable insights, its spatial resolution and precise boundaries are limited. These limitations stem from factors such as the residual momentum of the positron and the dispersion angle of the annihilation photons, resulting in less detailed depictions of the examined areas. Now, we will explain some important concepts related to PET/CT acquisition procedure:

- Positron range: As stated, the final acquired image represents the locations of annihilation photons, not the positron emission locations. These positrons can travel a short range before combining with a free electron and emitting two photons. This causes a misplacement of these events, compromising the

precision of the image.

- Noncollinearity of the annihilation photons: The residual momentum of the positron before it combines with an electron causes the angle between the two emitted photons to deviate from exactly 180°. This deviation results in a blurry image, with the extent of blurriness potentially being 1 or 2 mm, depending on whether the ring diameter is 50 or 90 cm[31].

- Detector size: An important factor affecting the spatial resolution of the final image is the spatial resolution of the detector itself, which is an intrinsic property of the equipment.

- Detector crystals: The materials used in the detector can affect decay time, high light output, and sensitivity.

- Positron emitting radiotracers used in PET: Numerous radiotracers have been developed for PET. The most well-known of these is ^{18}F-FDG (Fluorine-18 fluorodeoxyglucose), which offers advantages due to its relatively long half-life (110 minutes), allowing for transport to remote sites and extended study durations. Additionally, its low positron energy results in shorter travel distances for positrons, thereby combining with electrons more quickly and producing a more accurate and less blurry image. It can also be mass produced without a significant burden[31].

The major drawbacks of PET are the low spatial resolution and the absence of an anatomical reference frame, making it difficult to precisely pinpoint the location of tumors and the organs involved. This is particularly important in clinical studies where the accumulation of radiotracers must be determined as pathological or physiological.

Moreover, locating the site of pathological radiotracer accumulation helps with tumor staging and correct diagnosis. It has been a while ago since the incorporation of CT images with the PET has shown potentials to overcome this challenge and come up with a better tumor localization. This incorporation of two imaging modalities can take place in a manual fashion, where a reader compares the two images side by side, or through software that integrates them into one image. However, various factors like different scanner bed profiles, organ movement, and patient positioning may put an obstacle on the manual integration process. Therefore, scientists developed a special scanning device to acquire both image modalities simultaneously, overcoming these challenges and reducing cost and time.

One issue with PET is that some deeper parts of the body, surrounded by many organs, experience less photon annihilation because the positrons get absorbed or their energy dissipates in the body before combining with an electron. In contrast, in parts of the body like the skin, which is not surrounded by any organs, or the lung, which is mostly filled with air and has little tissue to slow down the positrons, the majority of the positrons react with electrons and consequently, there is no attenuation in the final image. As a result, not all parts of the body are represented at the same scale and quality in the PET image. To compensate, an attenuation map of the whole body, created with the help of CT scanners, is applied to the acquired PET image. This not only amplifies the attenuated regions in the body but also retains the image quality in areas that have not undergone attenuation[32].

1.3.2 Importance of automated segmentation in PET/CT

In recent years, PET/CT has become increasingly prevalent, leading to the development of more complex and advanced equipment, which in turn has resulted in higher data volumes and resolutions per scan. This has increased the workload for experts and requires more time and effort for the interpretation of the scans[33]. The improvement in image quality demands that readers spend more time on certain parts of the image, thereby increasing the risk of missing critical areas in the scans for nuclear medicine physicians and radiologists[34]. Developing expertise in PET/CT image reading and interpretation requires significant training and time. Additionally, the presence of multiple scanner types with varying settings introduces inconsistency across different systems. These factors contribute to the high susceptibility of reading scans and identifying tumor regions to inter- and intra-observer variability[35].

Automating tumor segmentation allows for the efficient identification of total tumor burden for treatment and prognostication and the usage of automatic quantitative biomarkers/radiomics, thereby assisting experts and providing support [36], [37], [34]. This automation in PET/CT significantly accelerates the segmentation of tumor regions, saving experts considerable time. Furthermore, with technological advancements, the outcomes of automated models are improving in consistency and robustness. Automated models, particularly those based on deep learning, are more objective and do not suffer from the subjectivity associated with human skill variance and fatigue.

1.4 Deep learning and its advances for imaging analysis

Deep learning uses multi-layered neural networks to automatically learn data representations, revolutionizing fields like speech, image recognition, and natural language processing[19]. These models vary in size, shape, architecture, and complexity. For years, convolutional neural networks (CNNs) had been the state-of-the-art for classification/segmentation tasks related to images and videos in the deep learning domain. They excelled in image/video segmentation tasks due to their ability to automatically learn hierarchical features from data. CNNs exploit spatial correlations within images through convolutional layers, capturing local patterns and gradually integrating them into higher-level representations. However, CNNs lack the innate ability to capture long-range dependencies and global context effectively[9].

More recently, transformer architectures[3] initially designed for sequential data, have gained traction for segmentation tasks. Transformers, especially vision transformers (ViTs)[4], offer global context understanding without relying on fixed-size convolutional kernels, which is crucial for capturing intricate spatial relationships in images. They leverage self-attention mechanisms to weigh the importance of different image regions adaptively. This adaptability, combined with efficient training and scalability, has led the transformer architectures progressively competing with CNNs in image segmentation tasks.

1.5 Problem statement

Given the advances in recent deep learning architectures for image segmentation, we adapt and evaluate three algorithms for lesion segmentation task in PET/CT images: one solely based on CNN, one as a hybrid method using a CNN and a transformer, and one completely based on transformers.

- **nnU-Net (no-new-U-Net)**: nnU-Net[38] is a self-configuring approach based on CNNs that can automatically configure itself by extracting relevant parameters for preprocessing, network architecture selection, training procedures, and postprocessing for the segmentation task.
- **Segment Anything Model (SAM) using nnU-Net box prompts**: SAM[8] is an open-source model trained using transformers used for segmenting any object, even without being trained on that class.
- **Swin-Unet**: Swin-Unet[9] is a pure transformer modeled after U-Net[39] with skip-connections to extract local-global semantic features from the image patches.

Our problem statement at its core is:
Performance analysis of three recent deep neural network-based automated approaches for tumor segmentation in whole body PET/CT images.

1.6 Organization of the thesis

The rest of this thesis is organized as follows:

Chapter 2 provides an in-depth discussion of the literature on tumor segmentation in PET/CT images. Chapter 3 describes the materials and methods in detail

for our algorithms. Chapter 4 reports the quantitative and visual results from our experiments. Chapter 5 discusses the results. Chapter 6 provides the limitations of our work. Finally, Chapter 7 concludes the thesis and provides future prospects for advancements in the current work.

Chapter 2

Literature Review and Related Concepts

In this chapter, we provide details of the search methodology we used to gather research articles from existing literature, and then elaborate on the studies focusing on tumor segmentation in PET/CT images in detail.

2.1 Searching methodology

We conducted our initial search across the most popular academic databases for paper extraction in September 2022, including: Google Scholar, IEEE Explore, PubMed, and ScienceDirect. The search terms considered are: artificial intelligence, deep learning, machine learning, biomedical image analysis, biomedical image segmentation, segmentation, PET/CT, and tumor delineation. We tried to extract the most relevant papers based on the following criteria: the number of citations, recency, and venue (journal/conference) of publication.

Among the relevant studies retrieved by our search, deep learning-based methods are predominant, however, we also discuss image processing/computer vision and machine learning-based methods in our review to highlight the differences of these with the deep learning-based methods.

We will now elaborate on the relevant studies in the literature, starting from older methods like image processing, and move on to the most recent ones.

2.2 Image processing and computer vision-based methods

Note: Throughout this and following chapters, we may use the terms 'image processing' and 'computer vision' interchangeably. We refer to the same concept by using either of these terms, although in the pre-deep learning era, they enjoyed distinct usage in the research community.

Image processing-based methods refer to techniques that utilize handcrafted rules, filters, and algorithms to extract features, and segment the targeted object. These features are mostly manually designed based on domain knowledge and heuristics[40]. Thresholding has been one of the most well-known approaches in the image processing realm. However, thresholding is prone to many errors and can be easily biased. For example, FDG uptake is indicated by high standard uptake value (SUV) in the chest for acute inflammatory and infectious diseases, such as pneumonia, whereas some tumors can have low SUV[41]. Moreover, there are many factors that lead to variability in PET images, including, but not limited to, different acquisition scanners, image reconstruction methods, calculations of SUV based on scanner type, and the

noise involved[42].

One example of this category of work is the study by D. Han et. al,[43]. They used MRFs for the segmentation problem. Different energy functions were developed for PET and CT modalities, and ultimately, the global optimal solution was derived by computing a single maximum flow. They reported an average Dice score (also called Dice Similarity Coefficient, discussed in Chapter 3) of 85% on 16 test images, and reported a 10% increase compared to graph cuts methods.

Image processing-based models can be effective, but with the emergence of new machine learning and deep learning models that outperform conventional methods, there has been a shift towards these algorithms in academic research. The reasons for this shift include but are not limited to:

- Sophisticated image processing models are computationally heavy and require extensive numerical calculation, making real-time inference a challenge. In contrast, deep learning models benefit from various optimization techniques and frameworks that significantly enhance inference speed on current graphics processing units (GPUs)[44].

- The availability of larger and higher-quality medical imaging datasets meets the demands of data-intensive deep learning models, facilitating significant improvements and enhanced performance in previous tasks.

- Deep learning and traditional machine learning methods show comparatively less susceptibility to dataset biases compared to image processing-based models, if trained and validated properly[45].

2.3 Traditional machine learning

Traditional machine learning (non-deep learning) refers to methods that classify, segment, or differentiate between distinct classes based on learning from handcrafted features extracted from images in a training dataset provided by experts. Usually, prior to the feature extraction, the images could undergo some pre-processing steps (e.g. resampling and zero-padding, resizing, augmentation, intensity normalization, windowing and leveling for CT, and conversion to SUV maps for PET[34]). These features can be of any format or length, numerical or categorical. We will provide examples of two such studies here.

In the work of K. Ikushima, et al.[46], a support vector machine (SVM) as the machine learning classifier was employed to delineate gross tumor volume (GTV) regions. The SVM learned image features around GTV contours determined by radiation oncologists during the training step. These features included voxel values and image gradient magnitudes from planning CT, PET, and diagnostic CT images. They compared the bare SVM model versus a model that applied an algorithm called optimum contour selection (OCS) on top of SVM outputs from their previous study[47]. The authors used images from 14 lung cancers patients for training and reported an average 3D Dice score of 77.7% using the proposed framework, and 50.7% for the OCS algorithm.

E. Grossiord et. al,[48] proposed an automated method for segmenting lymphoma lesions in 3D PET/CT images using hierarchical image models and machine learning. They used component trees to represent PET images and compute PET/CT descriptors for each region, followed by random forest classification to categorize nodes as

lesions, organs, or non-relevant structures. The method achieves a 92% lesion detection rate with mean Recall and Specificity of 73% and 99% respectively, without user interaction. However, there is a 35% volume overestimation compared to ground truth, suggesting the need for refining feature selection and incorporating anatomical atlases to address false positives and volume overestimation in future iterations.

There are many works done related to traditional machine learning, but because at its core, the segmentation problem could be handled much more conveniently using deep learning models, there are more recent research work published in that area.

2.4 Deep learning

In general, methodologies in deep learning that have manifested high potentials for image segmentation tasks, fall into two categories: CNN-based and Transformer-based.

2.4.1 CNN-based models

Definition of CNN models

CNNs are among the most successful architectures in deep neural networks. CNNs are particularly effective for grid-like data, such as images or videos. They are widely adopted in fields like object detection, image recognition, and image classification[49]. The learning process involves feeding training data to the network, computing outputs, calculating a loss function, and updating the network's parameters (weights and biases) based on a process called **backpropagation**[19]. During backpropagation, the loss function is differentiated with respect to each weight and bias in the

network's topology, allowing the parameters to adapt for better output predictions. This architecture is particularly useful in image and video domain because of a concept called kernels (filters). These kernels are arrays (1D, 2D, 3D) that are slid over the image and extract some patterns from it based on the weights present in the arrays. Each kernel could have a specific functionality, extracting unique features from the image. The output of applying a kernel forms a feature map, which emphasizes certain features in the image, such as edges, textures, or other patterns. During backpropagation, the kernels' weights get updated based on the losses calculated from the training set to adapt the network to learn features that are more suitable to the learning objective[50].

Favored CNN-based architectures for segmentation

CNNs have demonstrated considerable potential to compete with human-level expertise in image segmentation tasks with models like Mask R-CNN[51], SegNet[52], and DeepLab[53]. Another notable model is U-Net[39], which has shown remarkable results for image segmentation. The U-Net architecture is a fully convolutional neural network upon which our first and second segmentation approaches are based on. It comprises a contracting path, a bottleneck, and an expansive path. The contracting path captures context and extracts features from the input image. The bottleneck acts as a bridge between the contracting and expansive paths, reducing spatial dimensions while retaining essential information. The expansive path is responsible for generating the segmented output. Below are listed the salient components of a standard U-Net:

- **Contracting path**: This involves several convolutional layers that extract

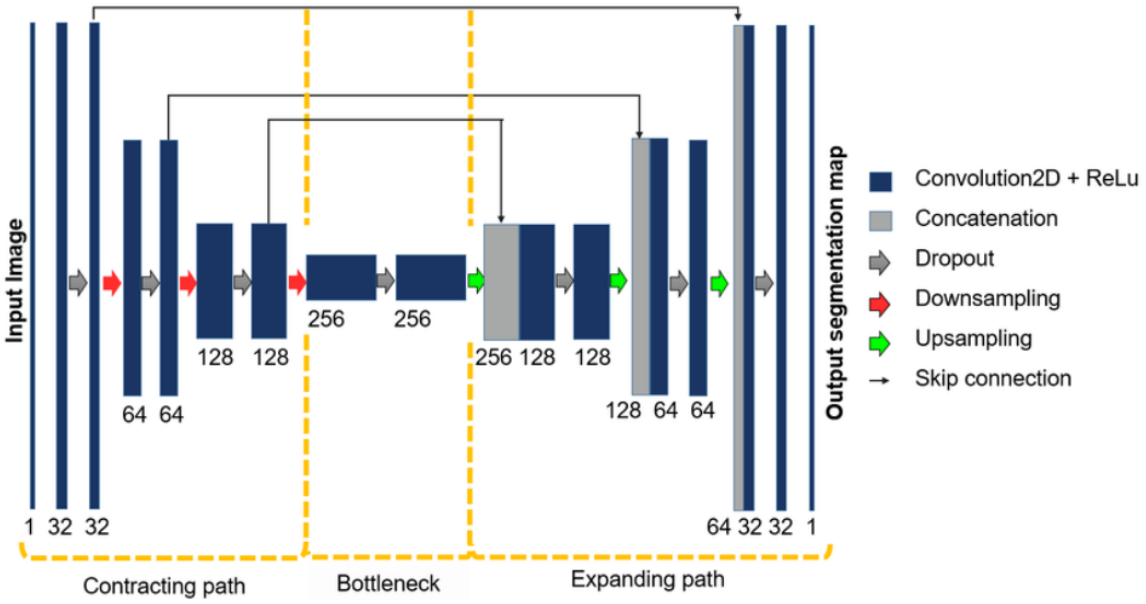

Figure 2.1: U-Net architecture. Adopted from[2]

features from the images, followed by rectified linear unit (ReLU)[54] activation functions and either average or max-pooling layers[55]. As the images pass through these layers, deeper and more complex features are extracted, and the dimension of the subsequent images (features maps) is further reduced.

- **Bottleneck**: This section between the contracting and expansive paths retains the most critical features of the image while further reducing spatial dimensions. It includes convolutional layers but omits pooling layers to prevent loss of important features. Dropout layers[56] may also be included to mitigate overfitting.

- **Expansive path**: Operating inversely to the contracting path, this path involves upsampling (via a process called transpose-convolution) the images back to their original spatial dimensions. During this process, spatial resolution is

enhanced by concatenating feature maps from the contracting path with the features after upsampling, ensuring both high resolution and the retention of deep features from the bottleneck layer. A ReLU activation function is applied after deconvolution layers to refine the features further.

- **Skip connections**: An innovative aspect of U-Net is the use of skip connections, which allows the network to bypass intermediate layers and connect outputs directly to deeper layers. This helps retain fine-grained details in the image and addresses the vanishing gradient problem.

- **Final layer**: This is a 1×1 convolutional layer with a sigmoid activation function, producing multiple channels that correspond to the different classes to be segmented. Each channel outputs a probability map indicating the likelihood of each pixel belonging to its respective class.

- **Loss Function**: The original implementation of U-Net used the cross-entropy loss[57], a common choice for segmentation tasks.

Related studies in tumor segmentation using CNN-based models

L. Xu et al.[58] employed V-Nets[59] (a 3D version of U-Net) and W-Nets[60] for bone lesion identification for diagnostic assessment of multiple myeloma (MM), using PET/CT modalities in conjunction. They conducted four experiments: V-Net with CT only, V-Net with PET only, V-Net with PET/CT, and W-Net for PET/CT. The last experiment yielded the highest scores in terms of Dice score (72.98%), Recall (73.5%), Precision (72.46%), and Specificity (99.59%). However, their dataset was

small, containing only 12 clinical PET/CT images, which may introduce bias. Additionally, they produced synthetic digital phantoms of ^{68}Ga-Pentixafor PET scans, which might not be ideal for testing in real-world scenarios.

X. Fu et al.[61] utilized two U-Nets for tumor segmentation: one network was solely trained on PET images to extract and highlight tumor regions. The probability map from this network's output was then fused with the decoder part of the second U-Net, which processed CT images. The rationale was that PET and CT modalities do not necessarily contain complementary information, hence, the study focused more on PET images. The alogorithm was tested in two datasets, one containing patients with non-small cell lung cancer (NSCLC) and one with soft tissue sarcomas (STSs)

Zhong et al.[62] trained two U-Nets separately on PET and CT images to segment lung cancer. Subsequently, a graph-cut co-segmentation algorithm was applied to the results to achieve a more consistent segmentation. However, this paper lacked detailed information and did not adequately compare their work with other studies. Additionally, their validation techniques were not mentioned properly and is vague.

Dirks et al.[63] developed a two-step localization and segmentation approach for tumor in malignant melanoma cancers, employing five fully connected networks; three trained on PETs with different spacings, and two on CTs with different spacings. A significant drawback of this study was the use of a thresholding technique to discard lesions smaller than 1 ml. Moreover, they calculated the SUV threshold value based on the SUV in the liver, which could be misleading due to tumor heterogeneity.

B. Huang et al.[64] utilized a deep neural network inspired by U-Net and processed two 512×512 PET and CT images to contour head and neck cancer tumors. They employed leave-one-out cross-validation and devised two datasets, one containing 17

images and the other 5. These sample sizes seem insufficient to train a robust and generalizable network. However, they reported an average Dice score of 74.1% over the two databases.

X. Zhao et al.[65] developed two V-Nets to separately extract features from PET and CT modalities. Subsequently, a fusion framework consisting of several cascaded convolutional layers was used to re-extract features from the two feature maps using weighted cross-entropy minimization. The algorithm was trained on a dataset including 84 patients with lung cancer, and the Dice score of 85% was reported.

Xiang et al.[66] utilized a dual-stream encoder to extract complementary information from PET and CT modalities, while a decoder was employed to fuse these distinct modality features to segment tumors in lung cancer. Additionally, two separate decoders were used to preserve modality-specific features of PET and CT images. Three conditional generative adversarial networks (GANs)[67] were utilized: one for PET, one for CT, and one for discriminating between the two modalities. The networks were trained to recognize and penalize discrepancies between the actual images and the segmentation outputs. At the end, average Dice scores of 75.52% and 77.72% were reported for the network trained on CT and PET images respectively.

Kumar et al.[68] emphasized that most papers often overlook the spatially varying visual characteristics that translate distinct information in each modality. They adopted two encoders, one for each modality, to extract modality-specific information. A co-learning scheme consisting of a CNN was then used to derive spatially varying fusion maps, followed by a fusion operation to weight different features differently. The scheme was trained and evaluated on a dataset of lung cancer patients, and they segmented lung, mediastinum, and the tumors inside these areas. A mean Dice score

of 63.85% was reported at the end.

The HECKTOR Challenge[69] focused on tumor segmentation in the head and neck using PET/CT images. The best performing method[70] (Dice score= 75.9% on test set) employed 3D U-Nets[71] with squeeze-and-excitation norm layers in the decoder section. The runner-up[72] implemented a 3D U-Net (derived from nnU-Net[38]) and a hybrid active contour for refinement on the test set only, based on the value of the normalized surface Dice (Dice score= 75.2% on test set).

S. Jemaa et al.,[73] proposed a method that utilizes a 2D U-Net for segmenting non-Hodgkin's lymphoma and advanced NSCLC from ^{18}F-FDG PET/CT images, focusing particularly on liver and lung detection through connected component analysis. This approach was applied to the head-neck, chest, and abdomen-pelvis regions. Additionally, a 3D U-Net was employed for the same purpose across these three regions. The final tumor mask was derived by averaging the results obtained from the 2D and 3D segmentation processes. An average 3D Dice score of 88.6% was reported on 1124 hold-out non-Hodgkin's lymphoma cancer scans, and 93% Recall on 274 NSCLC hold-out scans.

A. R. Groendah et al.[74] performed a comparison between four thresholding methods, six machine learning algorithms, and one deep learning-based model for tumor segmentation in patients with head and neck cancer. The thresholding methods were based on the absolute SUV, a percentage of the maximum SUV, 41% of the maximum SUV, or the Laplacian of Gaussian method (LoG). The machine learning algorithms included Gaussian naïve Bayes, linear discriminant analysis, quadratic discriminant analysis, logistic regression, linear support vector machines (SVM), and random forest (RF). The deep learning-based method, a 2D U-Net, was trained on three conditions:

PET only, CT only, and PET/CT together. Ultimately, the U-Net's results with PET/CT modalities and windowing outperformed the other methods (mean Dice score of 75%), demonstrating the superiority of deep learning approaches over traditional models.

Z. Zhong et al.[75] developed two 3D U-Nets with skip connections and shortcuts between the two networks to share parameters and help each other extract complementary information from the images. A weighted average of the two scores related to the two maps from the networks was calculated and reported as the final score. Their dataset included 60 patients with NSCLC who received stereotactic body radiation therapy. The average Dice score on network trained on CT modality was 86.1%, whereas for the network trained on PET, the score was 82.8%.

J. Xie and Y. Peng[76] enhanced a 3D nnU-Net with squeeze and excitation blocks[77] to boost meaningful features while suppressing weaker ones. Their results on the head and neck tumor segmentation challenge showed slight improvements over the original nnU-Net.

L. Li et al.[78] utilized 3D CT images and fed them to a fully convolutional network to obtain a probability map to distinguish between NSCLC tumor regions and surrounding tissues. Subsequently, a fuzzy variational model was proposed to incorporate the probability map from CT and the intensity of PET images to determine the tumor regions. Additionally, a split Bregman algorithm was applied to refine the boundaries. An average Dice score of 86% was reported on 36 scans held out for testing the algorithm.

L. Sibille et al.[79] utilized a cascaded methodology for tumor segmentation in PET/CT images. Initially, a stacked ensemble of 3D U-Net CNNs processed the

images at a fixed 6mm resolution. Following this, a refiner network, consisting of residual layers, improved the 6mm segmentation mask to achieve the original resolution. The approach consisted of two modules: they first analyzed global patterns and dependencies at a coarse level using an ensemble of modified U-Net networks, while the second refined the coarse segmentation using the original image. Training involved minimizing a composite loss function, including Dice loss, cross-entropy loss, and Sensitivity loss, with separate training for each model component. Additionally, post-processing involved exploring sequence-based models to reduce false positives, but these were ultimately removed due to similar results. This study uses a dataset of three cancers, namely, lung cancer, melanoma, and lymphoma (same as that used by us as described in Chapter 3). They had 84 volumes held out for testing and reported an average Dice score of 68% on this set. No cancer-specific analysis was conducted in this study.

The study by Y. Peng et. al,[80] proposes a false positive reduction network for tumor segmentation in the same dataset as ours, regardless of the cancer type. Initially, a self-supervised pre-trained global segmentation module roughly delineates candidate tumor regions using a pre-trained encoder. A local refinement module then refines these regions by removing false positives. The global segmentation module employs a ResNet50[81] encoder pre-trained via contrastive learning on concatenated PET/CT images, while the local refinement module uses a 2D U-Net with 5-channel input data. The Dice score on a hold-out set of 200 volumes was not exactly mentioned but according to the authors, they were placed among the top 7 scorers in the AutoPET 2022 challenge[82]. The best performing algorithm holds a Dice score of 62.26%.

Some novel approaches[83] integrated the idea of reinforcement learning into the

2D U-Net architecture, highlighting the importance of updating the network's weights through online learning to help the model continuously adapt to new images.

2.4.2 Transformer-based models

Introduced in 2017, transformers[3] are designed to handle sequential data, unlike CNNs which are predominantly used for image-related tasks. The core innovation of transformers is the **attention mechanism**, explained later. This technique is especially useful in tasks like language translation, where understanding the context provided by all parts of the sentence is crucial. While there are relatively fewer papers focused on this architecture, their numbers are increasing considerably. In Figure 2.2, the general architecture of a transformer is presented.

Vision Transformer (ViT)[4] is a variation of transformer (depicted in Figure 2.3) specifically designed to handle image data. The components of a ViT are as follows:

- **Input representation**: Unlike traditional CNNs that process raw pixel data directly, a ViT divides an image into fixed-size patches. Each patch is then flattened and linearly transformed into a higher dimensional space. In this space, each patch has been transformed into an "embedding". An embedding represents a patch as a vector of features, capturing its contextual information and visual details.

- **Positional encoding**: A number referred to as the positional encoding is concatenated to each image embedding to retain the positional information of each patch.

- **Self-attention mechanism**: This mechanism allows the model to learn the

contextual relationships among all image patches, and assign weights to all the elements of the input at the same time when predicting the semantic label. The representations along with the positional encodings are passed to this mechanism.

- **Multi-head attention**: By concatenating multiple attention mechanisms and running them in parallel, the model learns to determine the extent to which it should focus on each part of the image when predicting the segmentation map.
- **Feed-Forward networks (FFN)**: The output of the attention step is then fed to a feed-forward network (usually a normal multi-layer perceptron).
- **Upsampling**: Just like the idea of transpose-convolution in U-Net, feature maps in transformers experience a reduction in resolution due to patching and need to be upsampled to match the original image dimensions for pixel-wise classification. This is essential for the task of segmentation.
- **Segmentation head**: This is typically a convolution or a series of convolutional layers that process the upsampled features and predict each pixel's class.
- **Softmax activation**: This is applied across the channel dimensions of input image to obtain a probability distribution for each pixel, representing the probability of it belonging to each class in the set.

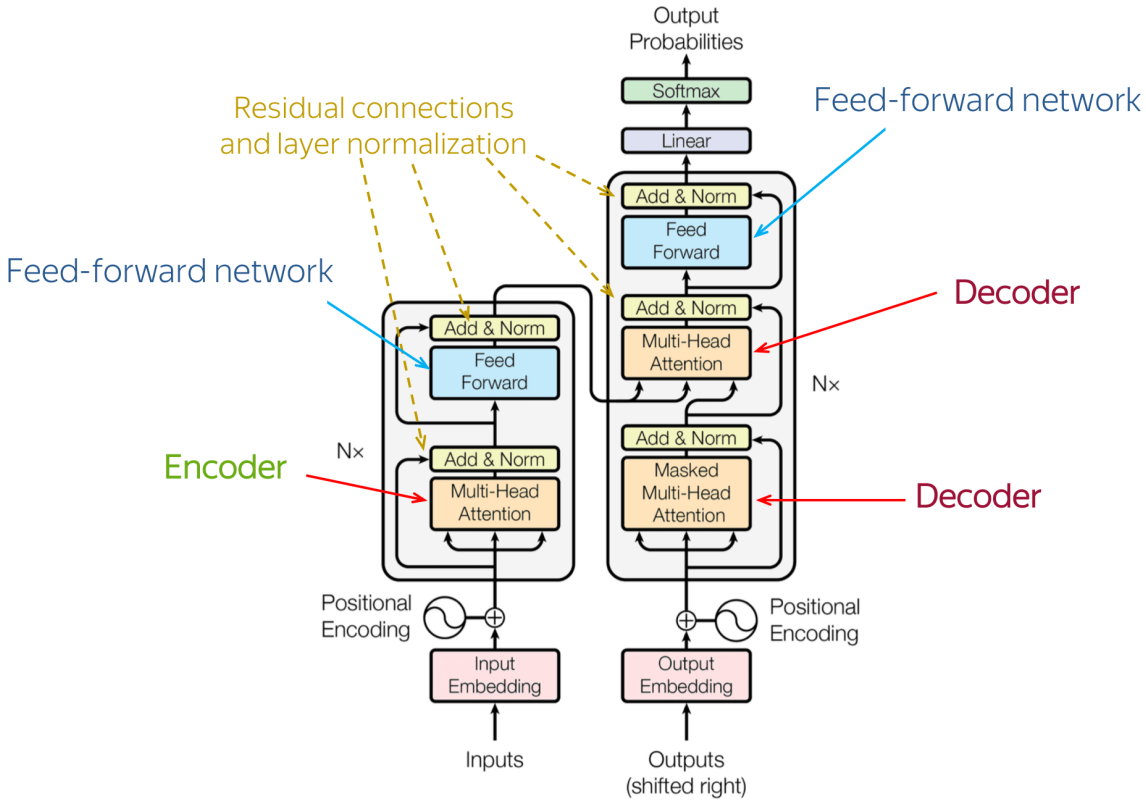

Figure 2.2: Transformer architecture. Adopted from[3]

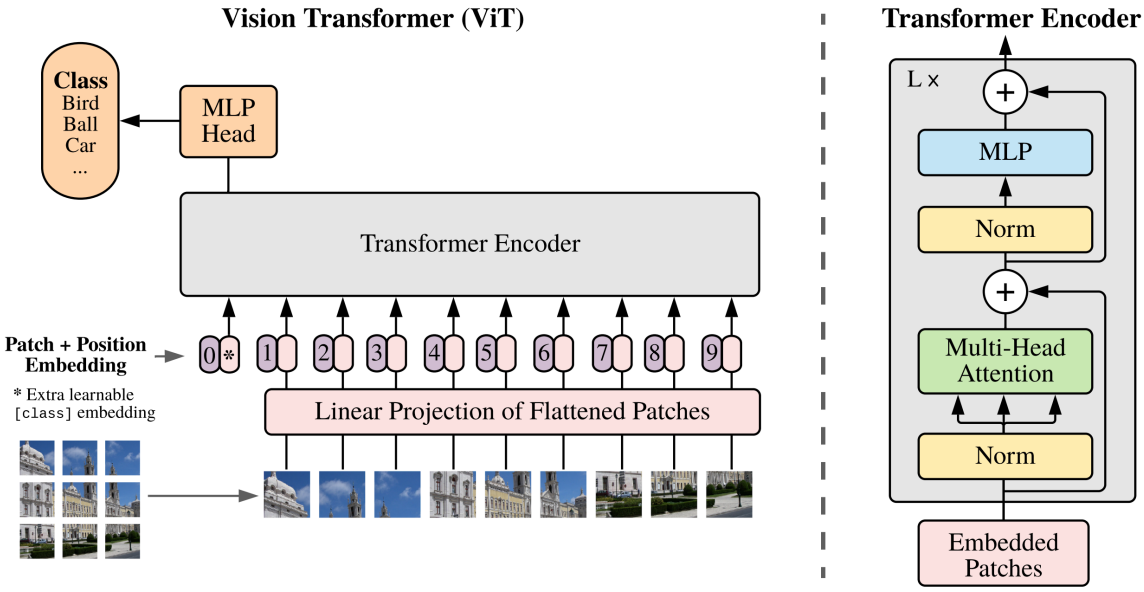

Figure 2.3: Vision transformer architecture for classification task. Adopted from[4]

Research articles related to transformer models or NLP-related architectures

Huang et al.[84] proposed a segmentation method for nasopharyngeal carcinoma (NPC) based on PET, CT, and parametric images (K_i images). They employed a generative adversarial network (GAN) with a modified version of U-Net and a transformer as the generator.

The method proposed by L. Huang et. al,[85] involves combining three consecutive slices from PET and CT scans along the depth dimension to form one volumetric representation for each modality. These representations are then concatenated as channels. Subsequently, an encoder-decoder pathway incorporating spatial and channel attention blocks generates a volumetric feature block representing the consecutive frames, alongside two types of slice-wise probabilities. A bidirectional Long

short-term memory (LSTM), also equipped with spatial and channel attention blocks, processes the volumetric feature block to integrate contextual information, ultimately generating three probability maps for the three slices.

One of the early studies addressing biomedical image segmentation problem using transformers is TransUnet[86]. The authors stated that a naive approach of using only transformers as the encoder of a neural network and then upsampling the hidden feature representations does not yield satisfactory results. This is due to the fact that transformers try to extract the global dependencies in any 1D sequence and this results in low-resolution features that lack the detailed local information. Hence, they took a hybrid CNN-transformer approach, where multiple levels of convolutional layers extract the local information from the image first. Then, the lowest-level features are fed to 12 layers of self-attentions. In the upsampling path, features extracted at each level of the CNN model are concatenated with the upsampled features from the transformer encoders. They compared their algorithm with V-Net, U-Net, DARR[87], and AttnUNet[88] on the Synapse multi-organ CT dataset[89] and reported better performance in the segmentation of various organs using TransUnet (average Dice score of 77.48%).

2.5 Drawbacks of current studies in PET/CT segmentation

- The majority of research studies have utilized CNNs for tumor segmentation tasks. There is a need for further investigation and exploration of transformer-based models in this domain.

- Several papers lack a detailed breakdown of their fold descriptions and do not adequately describe their approach to data splitting, which is crucial for mitigating any imbalance in the dataset.

- In our dataset (details of which will be elaborated in Chapter 3), there are both normal (no malignancy) and positive (with malignancy) cases. The positive cases include three types of cancer: lung cancer, melanoma, and lymphoma. Most studies utilizing this dataset do not provide performance comparisons of their models across these cancer types, an issue we addressed in our experiments.

- The majority of research studies have analyzed entire 3D or 2D images containing multiple tumors located at various body sites. To the best of our knowledge, no study has conducted an analysis of different tumor instances spread across different parts of the body. This is often due to the lack of distinct labeling for different tumor instances. Conducting such analyses would be pivotal for enhancing the automatic computation and analysis of quantitative imaging biomarkers. We addressed this gap in this thesis.

Chapter 3

Materials and Methods

In this chapter, we elaborate on the methodology followed by the implementation and evaluation of the three recent deep learning techniques for segmentation that we mentioned in Section 1.5. First, we describe the dataset used for training and evaluation. This is followed by an in-depth description of each technique, including details of its implementation for training (if necessary) and evaluation. Subsequently, we discuss the evaluation measures and statistical analyses conducted to obtain the results.

3.1 Dataset

We utilized a dataset provided by The Cancer Imaging Archive (TCIA)[90][29], which consists of whole-body PET/CT images with manually annotated lesions. This anonymized dataset was collected by University Hospital Tübingen, Germany, spanning from 2014 to 2018. It encompasses 900 patients (a total of 1014 image volumes), with some patients having undergone multiple visits. The cancer types represented

include lung cancer (168 patients, 168 image volumes), melanoma (177 patients, 188 image volumes), and lymphoma (144 patients, 145 image volumes). In total, 501 out of 1014 studies indicate malignant lesion, while the remaining 513 studies do not exhibit any PET-positive malignant lesions and are clinically normal. The distributions of the SUVs indicating PET-positive malignancy (presence of cancer) in this dataset in depicted in Figure A.9.

All image volumes were acquired using a single PET/CT scanner (Siemens Biograph mCT). The diagnostic CT scans (primarily covering the skull base to mid-thigh level) were conducted with intravenous contrast enhancement for most volumes, except for those patients with contraindications. The CT scan parameters are as follows:

- Reference dose of 200 milliampere-seconds (mAs)
- Tube voltage of 120 kilovolts (kV)
- Iterative reconstruction with a slice thickness of 2 - 3 mm

The whole-body PET/CT scans were performed one hour after the intravenous injection of 300-350 MBq ^{18}F-FDG. PET images were reconstructed using the ordered-subset expectation maximization (OSEM) method, which included 21 subsets and 2 iterations, along with a Gaussian kernel of 2 mm and a matrix size of 400×400.

All image volumes were reviewed jointly by a radiologist and a nuclear medicine physician to identify primary tumors and metastases. Following this identification, a radiologist with 10 years of experience in hybrid imaging segmented all FDG-avid tumor lesions (primary tumor and/or metastases, if present) using NORA image analysis platform, University of Freiburg, Germany. This segmentation resulted in ground truth PET/CT lesion masks. For each PET/CT image volume, regions containing

lesions were marked with an intensity value of 1 (foreground), while other regions in the image were marked with an intensity value of 0 (background) in the ground truth segmentation. Table 3.1 provides a summary of the dataset.

Table 3.1: Dataset summary

Variables	**Count**
Patients	900
Studies	1014
Normal studies	513
Positive/malignant studies	501
Positive slices	28280
Total slices	355147
Lung cancer studies	168
Lymphoma studies	145
Melanoma studies	188

3.2 Data pre-processing

3.2.1 PET/CT pre-processing

CT images were resampled to match the resolution of the PET images. All PET images were converted to SUV maps. The CT images, SUV maps, and ground truth lesion annotations were saved as 'nifti' files. These procedures were conducted in accordance with the protocols provided by the dataset authors[91]. We use the term 'PET' throughout this thesis, however, we refer to the usage of SUV maps for implementing and evaluating our models.

3.2.2 Ground truth pre-processing

The lesions in the ground truth were not confined to any specific organ or region in the body, even within the same cancer type. Consequently, a 3D-connected component analysis was performed to algorithmically identify connected neighboring voxels in lesions using the *connectedcomponent* function from the *sitk* package[92]. This analysis enabled the assignment of separate labels to connected lesion regions or components. For instance, if seven connected lesion components were identified, they were labeled sequentially from 1 to 7, with 1 corresponding to the component with the highest volume and 7 to the component with the lowest volume. An example of connected components in a single slice is illustrated in Figure 3.1. This connected component analysis, applied to both ground truth and predicted segmentations, facilitated experiments using SAM and allowed for lesion component-wise evaluation of the segmentation models' performances.

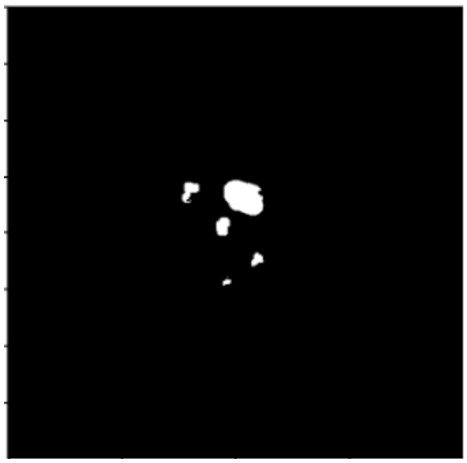

Figure 3.1: A slice from a segmentation volume map, containing five 2D connected components

3.3 nnU-Net

nnU-Net[38] is a neural network architecture developed for semantic segmentation tasks in medical image analysis. It extends the U-Net architecture, widely used for biomedical image segmentation, by incorporating advancements in deep learning and addressing specific challenges in medical imaging. The various components of this architecture are as follows:

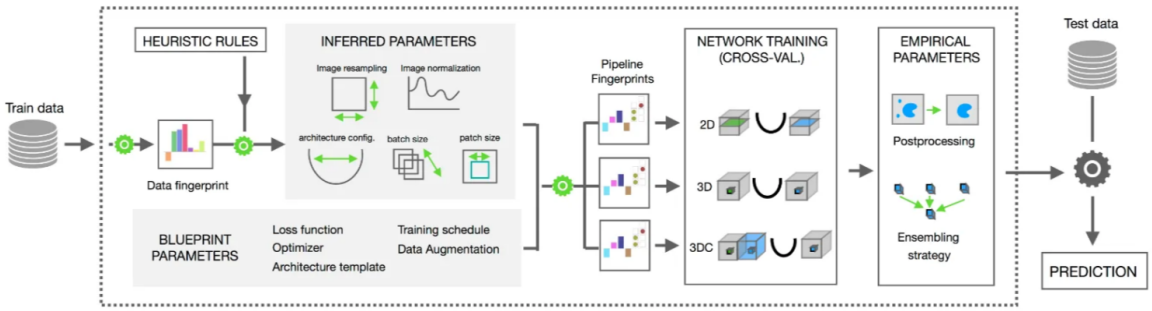

Figure 3.2: nnU-Net pipeline. Adopted from[5]

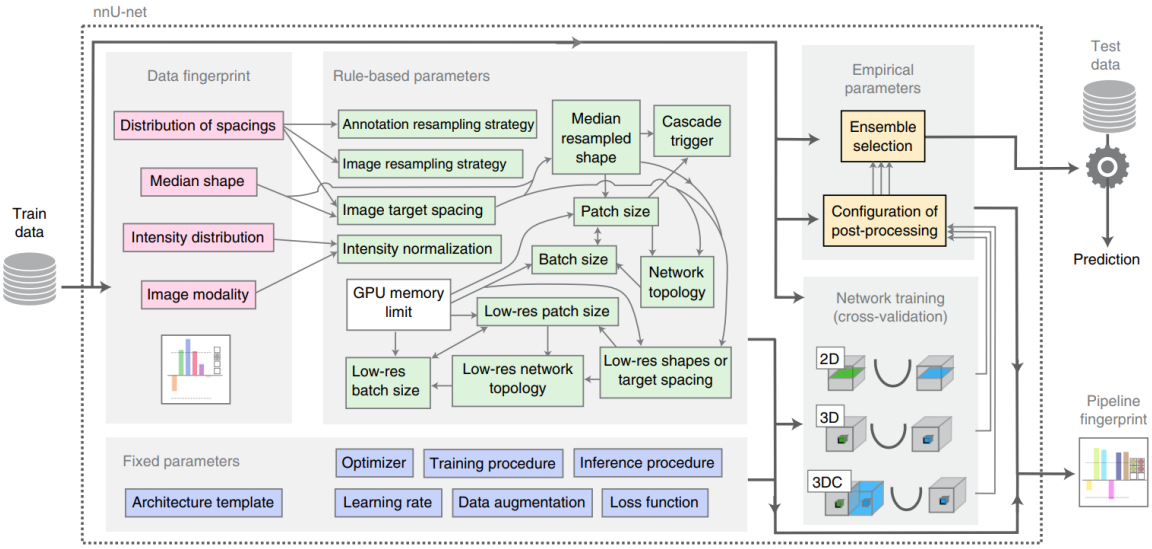

Figure 3.3: nnU-Net development procedure. Adopted from[5]

- **Architecture:** nnU-Net builds upon the U-Net architecture, detailed in Section 2.4. This pipeline evaluates three model types: a 2D U-Net, a 3D full resolution U-Net, and a cascaded 3D U-Net that first operates on low-resolution images before refining the segmentation maps using full resolution images.

- **Modular design:** nnU-Net introduces a modular design that allows for flexibility in configuring the network architecture based on dataset and task requirements. It features various modifications and enhancements to the original U-Net architecture, with adjustable parameters like batch size, patch size, and network topology to ensure an effective receptive field size, preventing the loss of contextual information.

- **Multi-Scale feature integration:** nnU-Net integrates multi-scale features from different network levels, aiding in capturing both local and global context, which is crucial for accurate segmentation in medical images where structures vary significantly in size.

- **Data augmentation and preprocessing:** nnU-Net employs advanced data augmentation techniques to enhance model robustness and improve generalization. This is particularly vital in medical imaging where datasets may be limited and highly imbalanced. Examples of augmentations include scaling, Gaussian blur, rotation, Gaussian noise, and variations in brightness and contrast.

- **Loss function:** nnU-Net typically uses a specialized loss function tailored for medical image segmentation, accommodating class imbalance, spatial proximity, and class-specific characteristics of medical structures. This function combines Dice loss and cross-entropy loss.

- **Training strategies:** nnU-Net utilizes specific training strategies such as progressive resizing, where the input resolution is gradually increased during training to help the network learn features at different scales and enhance its adaptability to images of varying resolutions.

- **Ensemble learning:** Some variants of nnU-Net apply ensemble learning techniques to boost performance. Multiple models trained with different initializations or architectures are combined for predictions, leading to enhanced segmentation accuracy and robustness.

- **Application in medical imaging:** nnU-Net is primarily applied to medical imaging tasks such as organ segmentation, tumor detection, and anomaly localization. Its precise delineation of regions of interest from medical images is crucial for treatment planning, disease diagnosis, and monitoring.

nnU-Net distills domain knowledge into three groups of parameters: fixed, rule-based, and empirical.

- **Fixed parameters:** These parameters do not require adaptation between datasets, such as architecture template, optimizer, learning rate, loss function, and data augmentation. They are set by the developer at the beginning of the experiments, and the model's performance ultimately depends on these choices.

- **Rule-based parameters:** The authors of nnU-Net have set heuristic rules in the form of explicit dependencies to calculate certain design choices (like patch size, batch size, and network configuration) based on the dataset fingerprint. The dataset fingerprint refers to a collection of characteristics captured from

the dataset, such as image size, voxel spacing information, class ratios, intensity distribution, and median shape. Examples of these rule-based decisions include:

  - *Image resampling strategy:* For anisotropic images, resampling in-plane is done with third-order spline and out-of-plane with nearest neighbor interpolation. For isotropic images, a third-order spline is used.
  - *Intensity normalization:* For CT images, global dataset percentile clipping and z-score normalization using the global foreground mean and standard deviation are performed. For PET, z-score normalization with per-image mean and standard deviation is applied.

- **Empirical parameters:** These parameters are adjusted at the end of the training procedure and include configurations for the post-processing step and ensemble selection. Non-maximum suppression is applied for post-processing, treating all foreground classes as one and retaining only the largest component. If cross-validation performance improves, this post-processing step is implemented, focusing on the largest component for each individual class. For ensemble selection, the model variant (2D, 3D, or cascaded 3D) that yields the highest cross-validation performance is chosen as the final model.

3.3.1 nnU-Net implementation

As discussed, fixed parameters are a subset of parameters in nnU-Net that are not optimized and are chosen based on the developer's decision, unlike the other two parameter groups (rule-based and empirical). Changing these parameters and tracking the extent to which the results of the segmentation differ is completely dependent

on the task at hand. Therefore, we decided to choose the loss function as a variable to tweak and observe how significantly the outputs deviate based on its choice. In addition to the loss function, we also experimented with the effects of various imaging modalities used as input. Below are the details of these variational changes:

1. **Loss function:** We tested five different loss functions to determine if model performance depends on this choice: Dice loss[93], Cross Entropy (CE) loss[94], Focal loss[95], a combination of Dice loss and CE loss, and Tversky loss[96].

2. **Modality:** We experimented with three different modalities: PET only, CT only, and PET/CT combined.

Initially, we conducted these experiments using 30% of the entire dataset as a pilot study to observe any significant deviations in the results by adjusting the aforementioned parameters. After identifying the most effective loss function and modality combination based on the highest scores from this subset of the dataset, we applied these settings to conduct a more comprehensive analysis on the entire dataset. Below are the formulas for the loss functions in a binary pixel classification scenario:

CE loss:

$$CE(p, y) = -y \log(p) - (1 - y) \log(1 - p) \tag{3.3.1}$$

$\boldsymbol{p}$ = predicted probability of a pixel belonging to the positive class

$\boldsymbol{y}$ = true label of the pixel

Focal loss:

$$FL(p, \alpha, \gamma, y) = -\alpha y(1-p)^{\gamma}\log(p) - (1-\alpha)(1-y)p^{\gamma}\log(1-p) \tag{3.3.2}$$

$\boldsymbol{p}$ = predicted probability of a pixel belonging to the positive class

$\boldsymbol{\gamma}$: a tunable focusing parameter to adjust the rate at which easy examples are down-weighted

$\boldsymbol{\alpha}$: weighting factor

$\boldsymbol{y}$: true label of a pixel

For any value of $\gamma > 0$, the misclassified samples can be controlled to be penalized more for hard samples, adapting the network to learn better representations for these data. For $\gamma = 0$, Focal loss is the same as α-balanced CE loss. The parameter α is used to address the class imbalance problem. In our experiments, we set $\gamma = 2$ and $\alpha = 0.5$.

Dice loss:

$$D_L = \frac{FP + FN}{2TP + FP + FN} \tag{3.3.3}$$

TP = True positive , FP = False positive , FN = False negative

Tversky loss:

$$T_L = \frac{2TP}{2TP + \alpha FP + \beta FN} \tag{3.3.4}$$

TP = True positive , FP = False positive , FN = False negative

($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) = Hyperparameters that control the balance between false positives and false negatives

Tversky loss is similar to Dice loss but addresses the class imbalance problem by weighting false predictions differently. For a higher Recall, normally a higher value of β is adopted, whereas for higher Precision, higher α is chosen[97]. In our experiments,

we set $\alpha = 0.3$ and $\beta = 0.7$.

Following extensive experiments documented in Appendix A for our pilot study, we determined that the Focal loss with the PET/CT combination yielded the highest Dice scores. Consequently, we proceeded with these parameters for our broader analysis of the entire dataset.

3.3.2 nnU-Net model implementation and evaluation on the entire dataset

The nnU-Net model was tested under the following conditions:

- No data augmentation was applied.
- Image spacing was adjusted to isotropic $[2, 2, 2]mm$.
- Validation was performed using 5-fold cross-validation.
- Focal loss (this loss function yielded the highest score among other loss functions as illustrated in Appendix A)
- PET images in conjunction with CT images were used (as this combination yielded the highest score in Appendix A)
- Training was conducted for 1000 epochs in each fold to obtain the fold-specific final model.

3.4 Segment Anything Model (SAM) with nnU-Net prompts

3.4.1 SAM

SAM[8], from the Segment Anything (SA) project by Meta Inc. (Figure 3.7), is a foundation model for image segmentation that aims to generalize image segmentation task towards a variety of images and pre-conditions (also called prompts). Foundation models are based on transformer architecture and are trained on a vast, diverse, rich, and versatile data that can be adapted to perform segmentation, classification, and other tasks[98], which led to significant advancements in the Natural Language Processing (NLP) domain, initially. These models are characterized to serve as a base model upon which specialized capabilities can be established. Examples of these architectures are BERT[99], GPT-3[100], and DALL-E[101], which are specialized in domains such as language, vision, etc.

SAM was trained on a dataset (called SA-1B) which contains over 1 billion masks on 11 million images, making it a powerful and large foundation model able to perform segmentation tasks using prompts. The different components of SAM are explained as follows:

- **Task Definition (Promptable Segmentation)**:
  - The task involves generating valid segmentation masks from any segmentation prompt, facilitating zero-shot generalization. Prompts are task-specific instructions designed to guide the model towards a desired output[102]. In the case of computer vision and image segmentation, this prompt could

be in many forms. In the case of SAM, box, point, and text prompts are supported. In Figures 3.5 and 3.6, examples of box and point prompts are presented, respectively.

  - The model ensures that at least one resulting mask is meaningful, even when the prompt is ambiguous.

- **Model Architecture**:

  - SAM comprises an image encoder, a prompt encoder, and a mask decoder.
  - The image encoder processes high-resolution inputs to produce image embeddings.
  - The prompt encoder handles various types of prompts, embedding them into a unified representation.
  - The mask decoder generates segmentation masks based on combined image and prompt embeddings, efficiently computing masks in real-time.

3.4.2 SAM implementation

We considered providing box prompts to SAM as it was shown to be more effective than providing point prompts[103]. These prompts are rectangle boxes surrounding the lesion areas from the predicted segmentations by the nnU-Net.

The nnU-Net algorithm tends to undersegment the tumor regions. To address this issue, we conducted an experiment where we first dilated the 3D lesion components obtained through connected component analysis on the predicted segmentation masks from the trained nnU-Net algorithm. This dilation was performed along the x, y, and z axes using a random pixel expansion ranging from 1 to 5. Subsequently, we enclosed

the tumor regions in 2D slices with bounding boxes and input these into the SAM algorithm, as SAM accepts 2D images. Examples of the dilation process are depicted in Figure 3.4. The code for the dilation and connected component analysis is provided in Appendix B.

This approach was tested under these variations:

- SAM was initially pre-trained using colored natural images; therefore, it takes input 2D images in three channels. Different permutations were analyzed for possible channel combinations: 3 PET, 2 PET and 1 CT, and 1 PET and 2 CT. The purpose of this test was to determine the extent to which each modality contributes towards improved segmentation outcomes.

- Box prompts from both dilated and non-dilated lesion components were tested. In total, three channel combinations along with two versions for both the dilated and non-dilated formats, resulted in $3 \times 2 = 6$ different testing configurations.

Figure 3.4: Comparison of dilated segmentation map (left) versus the original segmentation map (right). Adopted from[6]

Here are the inference settings for SAM:

- The biggest image encoder was utilized (ViT-H) with 636M parameters which outputs higher quality image representations.
- Contrast stretching was applied on the SUV images. The maximum SUV was mapped to 255, and the image data type was converted to unsigned integer with 8 bits (code available in Appendix B).
- Windowing and normalization were performed as pre-processing steps for CT images. A window level of 40 and a window width of 400 was selected[28]. After clipping the CT image values, they were normalized between the values of $0 - 255$ and data type was converted to unsigned integer with 8 bits (code available in Appendix B)[28].

- The model was run in inference mode using onnx (Open Neural Network Exchange) format[104]. The prompt encoder and mask decoder of SAM are lightweight, allowing it to run on any platform that support onnx-runtime. Onnx models are scalable, platform- and framework-independent, and fast for deployment in various applications.

The ONNX model incorporates the original SAM architecture, which includes an image encoder, a prompt encoder, and a mask decoder. Both the prompt encoder and mask decoder are lightweight and capable of fast inference. Throughout this thesis, we use this model in the "SAM with nnU-Net prompts" method for inference.

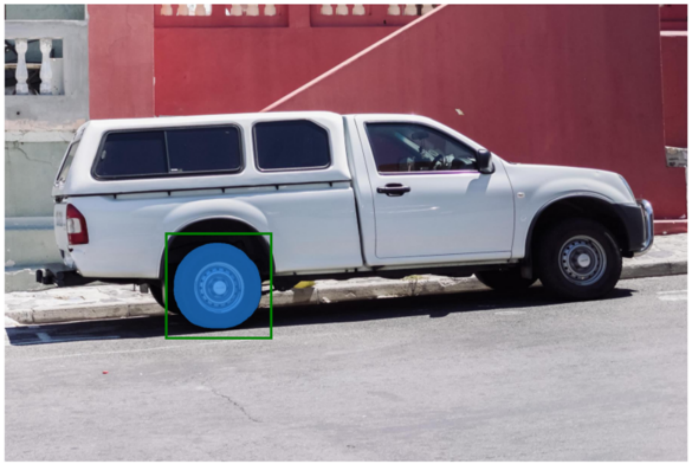

Figure 3.5: Box prompting in SAM. Adopted from[7]

Figure 3.6: Point prompting in SAM. Adopted from[7]

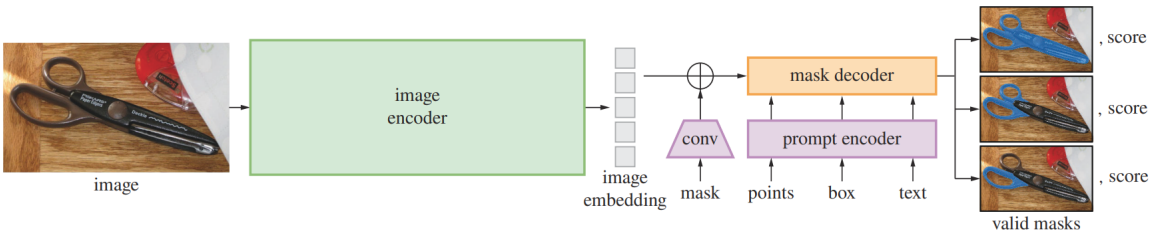

Figure 3.7: SAM architecture. Adopted from[8]

3.5 Swin-Unet

Swin-Unet[9] integrates the Swin Transformer architecture[105], tailored for vision tasks, with the skip connection strategy from the U-net architecture (Figure 3.8). It outperformed many other architectures (CNN-based or transformer-based) on the organ segmentation task on Synapse multi-organ CT dataset[105]. Here are the key components of the Swin Transformer:

- **Shifted windowing**: The Swin Transformer uses non-overlapping local windows that shift across layers, reducing computation compared to traditional transformers that apply global self-attention across all patches.
- **Hierarchical structure**: Similar to CNNs, it processes features at multiple scales, beneficial for capturing details necessary for image segmentation across different scales.

In Swin-Unet, traditional convolutional layers are replaced with Swin Transformer blocks that function as both encoder and decoder:

- As an encoder, it efficiently captures local features from the input image using shifted windows.
- As a decoder, it upsamples and refines the segmentation maps.

3.5.1 Swin-Unet implementation

Swin-Unet operates on two-dimensional (2D) images. We first trained the Swin-Unet using the entire dataset (including 355,147 slices) containing 28,280 positive slices (almost 8% of the entire dataset) under 5-fold cross-validation. However, the results

were not satisfactory enough as the model tended to flag the majority of cases as normal. Hence, we decided to increase the proportion of positive slices in the training set so that the model observes more malignant slices. Therefore, we selected all positive slices from the volumes and added normal slices at a ratio of 4:1 to the number of positive slices, chosen randomly from the rest of slices in each volume. If the calculated number of normal slices exceeded the total number of slices in a volume, all the remaining slices in that specific volume was taken. Ultimately, we compiled approximately 111,394 slices for training the model. It should be noted that only PET-positive volumes were utilized for training under 5-fold cross-validation, and no normal volumes were included during either the training or the internal validation phases.

The Swin-Unet model was implemented and tested under the following conditions:

- A model pre-trained on ImageNet dataset was used[106]. This helps the model find the optimal parameters for the objective task faster and more efficiently.
- Only PET images were used.
- The batch size is set to 32.
- All images were resized to $[224, 224]$ using nearest-neighbor interpolation.
- The loss function is a combination of Cross-Entropy (CE) and Dice loss[9].
- Stratified-5-fold cross-validation was conducted. This strategy ensures that there is a balanced number of positive slices in each fold.

- Within each fold, 80% of the patients among the 513 PET-positive image volumes was allocated for development (70% for training, 10% for internal validation), and 20% for testing.

- The test set features unique patients, ensuring no overlap with the training or internal validation sets. However, there were slices present in the training and internal validation sets which belonged to the same patient.

- All volumes used in the development part of this study showed PET-positive indications. Upon completion, we saved and evaluated the performance of the five models resulting from each fold. We also tested the normal volumes using one of the trained models obtained from a cross-validation-fold.

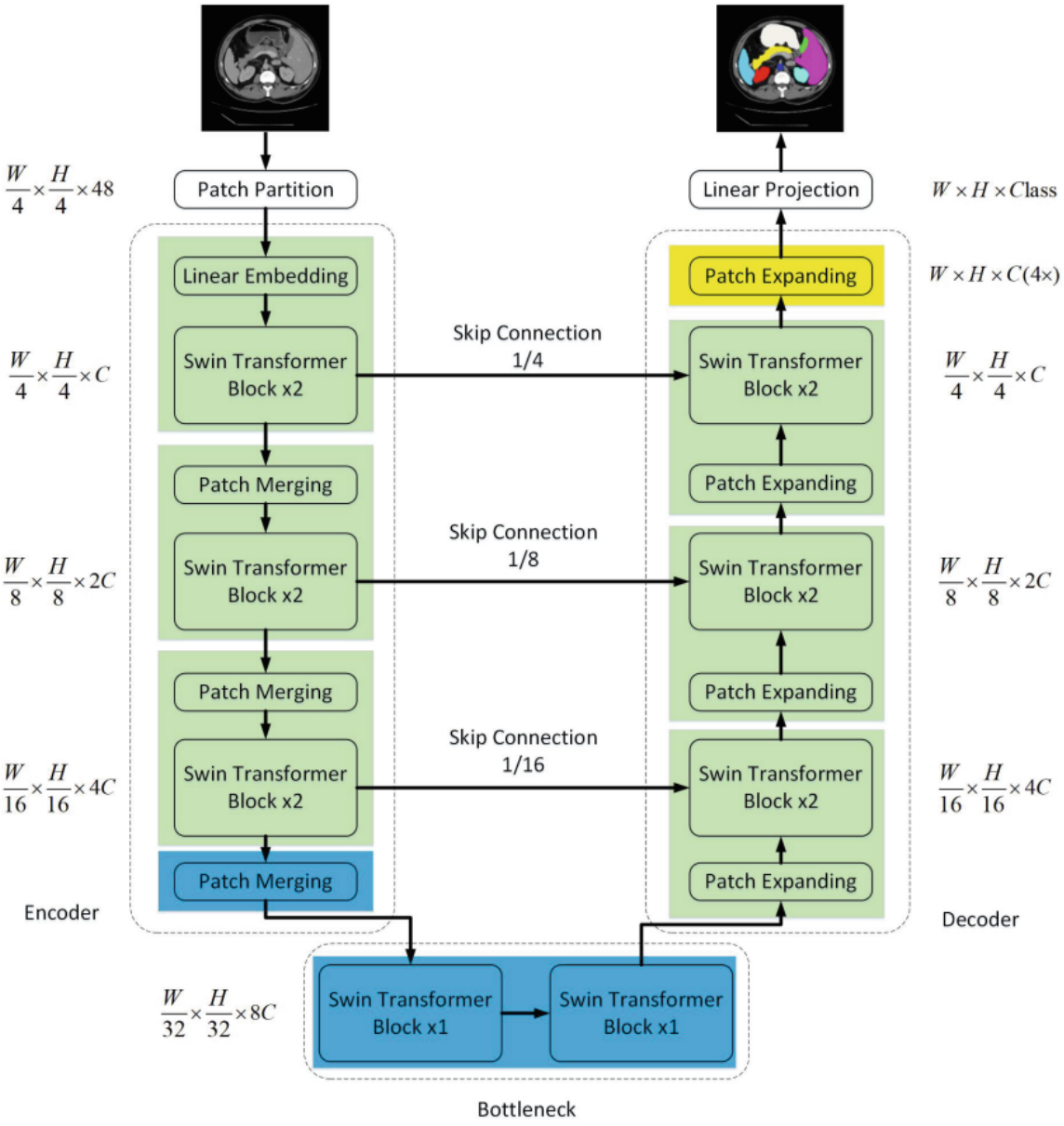

Figure 3.8: Swin-Unet architecture. Adopted from[9]

3.6 Evaluation metrics

To assess our methods, we used several crucial metrics: Dice score, Precision, Sensitivity (Recall), False Negative Rate (FNR), and Specificity. These metrics are computed at the volume level, not slice level.

The formulas for these metrics are:

$$\text{Dice Score} = \frac{2TP}{2TP + FP + FN} \tag{3.6.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.6.2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.6.3}$$

$$\text{FNR} = \frac{FN}{FN + TP} \tag{3.6.4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.6.5}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. A "True" scenario refers to correctly identified pixels in the prediction map: *True Positive* for correctly identified cancer pixels and *True Negative* for normal pixels. Conversely, a "False" scenario refers to inaccurately predicted pixels: *False Negative* for cancer pixels misidentified as normal, and *False Positive* for normal pixels incorrectly flagged as cancerous.

Spearman correlation coefficient is another measurement reported which calculates the strength of correlation between two ranked variables.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.6.6}$$

Where:

- d_i is the difference between the ranks of corresponding values x_i and y_i.
- n is the number of observations.

If the ranks of two variables are identical, the Spearman coefficient will be +1, indicating a perfect monotonic relationship. This means that both variables consistently increase or decrease together, following a monotonic function.

For all algorithms, we ensured to carry out the proper inference using the corresponding model to which the image volume of the test set belongs. For Swin-Unet specifically, in terms of normal volumes, we only chose the trained model from the fourth CV fold as it demonstrated the best performance in identifying normal cases.

3.7 Analysis paradigm

We performed two categories of analyses: volume-level and lesion component-level. The volume-level analysis involves comparing the predicted segmentations with the provided ground truth (considering all lesions as foreground) and computing relevant metrics for comparison. The lesion component-level analysis computes metrics for different lesion components (using the labels provided by the connected-component analysis) in each volume.

The first approach predominates in most research articles in the literature. However, the second type of assessment is not well-explored or detailed. This type of analysis is particularly useful for automated imaging biomarker analysis.

Another aspect of our analysis is evaluating the performance of various algorithms concerning different types of cancer. This perspective is often overlooked in the literature, but we aim to identify any significant and meaningful patterns in our results in this regard.

All segmentation models identify the lesions as foreground, marked by a single label with an intensity of 1. These are considered predicted positives, while the rest are considered predicted negatives. These classifications are used to calculate the metrics (3.6.1 - 3.6.5) as part of our volume-wise analysis for each cancer type and all cancers combined. An average of all the metrics over the number of volumes for which the scores are real numbers (not NaNs) are reported. Additionally, we note the missing rate (the percentage of cancer image volumes and cancer pixels missed).

We closely examine the misclassification of normal image volumes as having positive predictions for pixels. We also analyze the segmentation performance (using Dice scores) against the total lesion spread, i.e., the overall tumor volume in an image volume. Additionally, we provide visual examples of the predicted segmentation against the ground truth in patients with the highest overall tumor volume, using the slice with the highest tumor volume.

We applied the same connected component analysis to the predicted segmentation volumes to identify different lesion components within the predictions. We tabulated the number of 3D lesions per image volume in the ground truth and the predicted segmentations across all evaluated models (nnU-Net, "SAM with nnU-Net prompts" for different channel combinations, and Swin-Unet). We calculated Spearman's rank correlation for the number of lesion components in the ground truth versus the predictions for each of the three cancer types and for all cancers combined.

Dice scores were calculated between each pair of a ground truth lesion component and a predicted lesion component. The average and largest of these lesion component-wise Dice scores was obtained for each of the three cancer types and for all cancers combined. Then, the median and the range of these average and largest lesion component-wise Dice scores were calculated for each cancer type and for all cancers together. These scores were also compared with the median volume-wise Dice scores. We also captured the percentage of the highest Dice scores stemming from the largest lesion in the ground truth volume.

Chapter 4

Results

4.1 Volume-wise analysis

4.1.1 Cancer type-specific performance

As depicted in Figures 4.1, 4.2, and 4.3, 4.4, the nnU-Net algorithm consistently outperforms the other two approaches with respect to all performance metrics, across all three cancer types.

Furthermore, we noted that the fewer PET channels we used in the "SAM with nnU-Net prompts" experiment, the lower the scores became. This illustrates that PET modality is the primary modality (between PET and CT) contributing to tumor delineation and our initial hypothesis on using CT modality to provide the algorithms with skeleton details and help with a more precise boundary, fails for SAM. Additionally, dilation improves all the scores except Precision; the reason is by dilation we are capturing a larger area and this can potentially increase Dice and Recall scores, however, we are more likely to flag normal tissues as having cancer resulting in the

decrease of Precision. Also, dilation decreases FNR which is accomplished mainly due to the increase in the number of true positives.

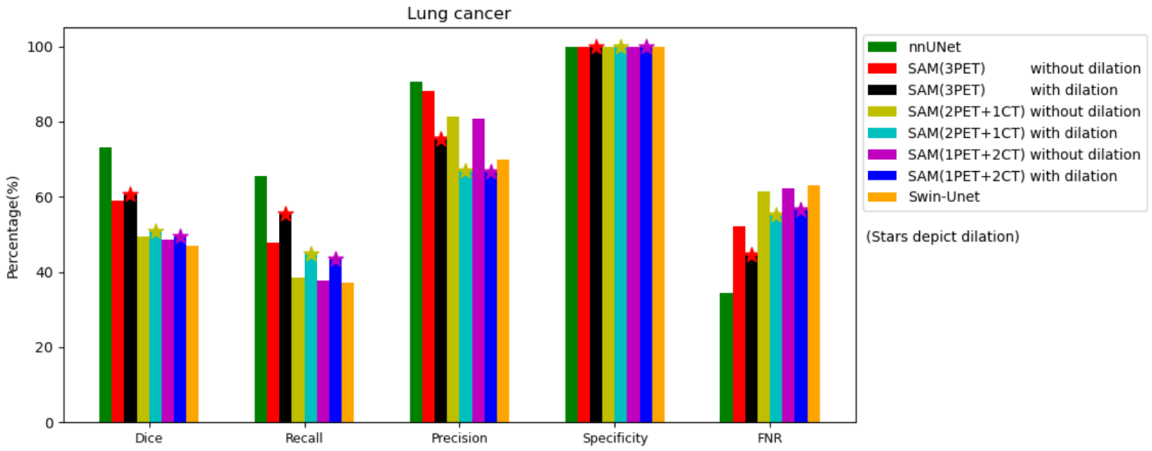

Figure 4.1: Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on lung cancer patients using five performance measures. 'Dice' denotes the Dice score.

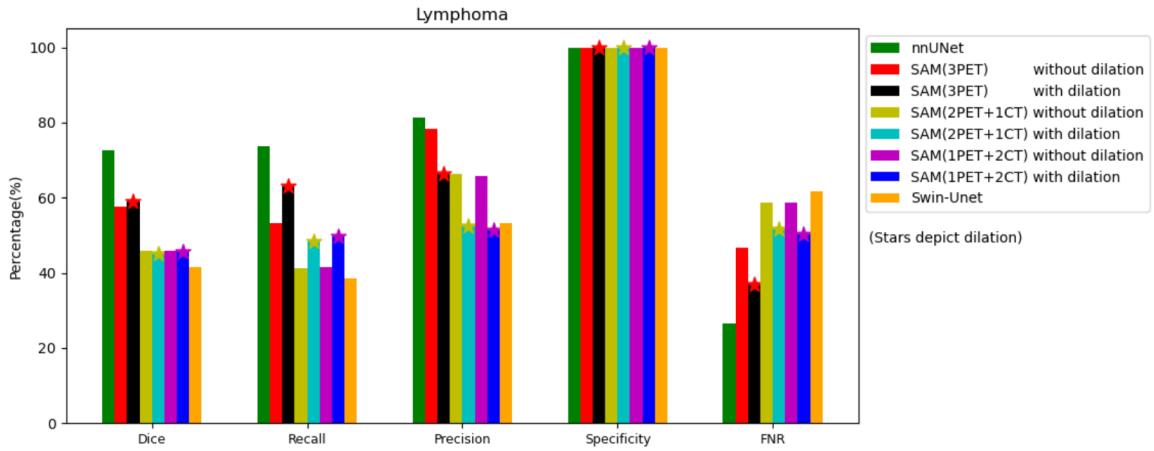

Figure 4.2: Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on lymphoma patients using five performance measures. 'Dice' denotes the Dice score.

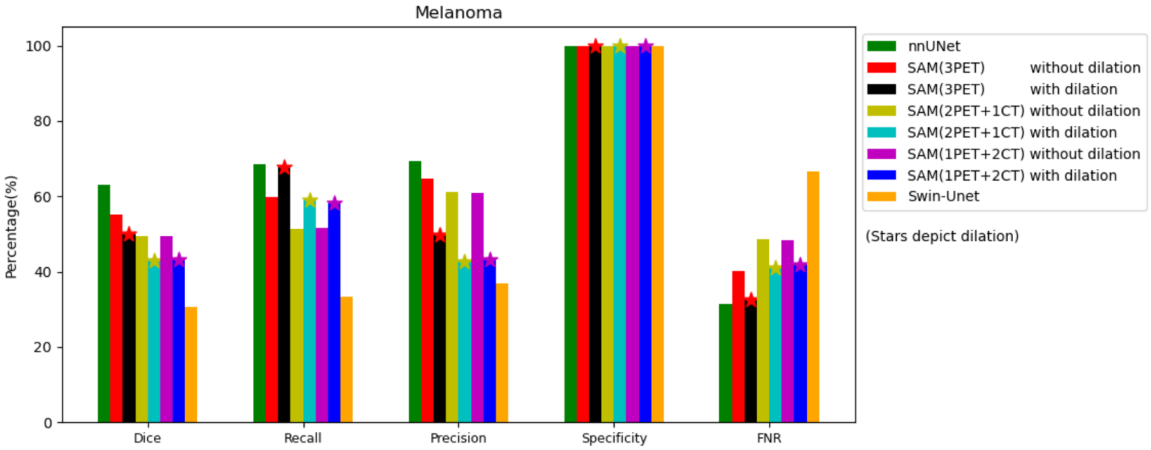

Figure 4.3: Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on melanoma patients using five performance measures. 'Dice' denotes the Dice score.

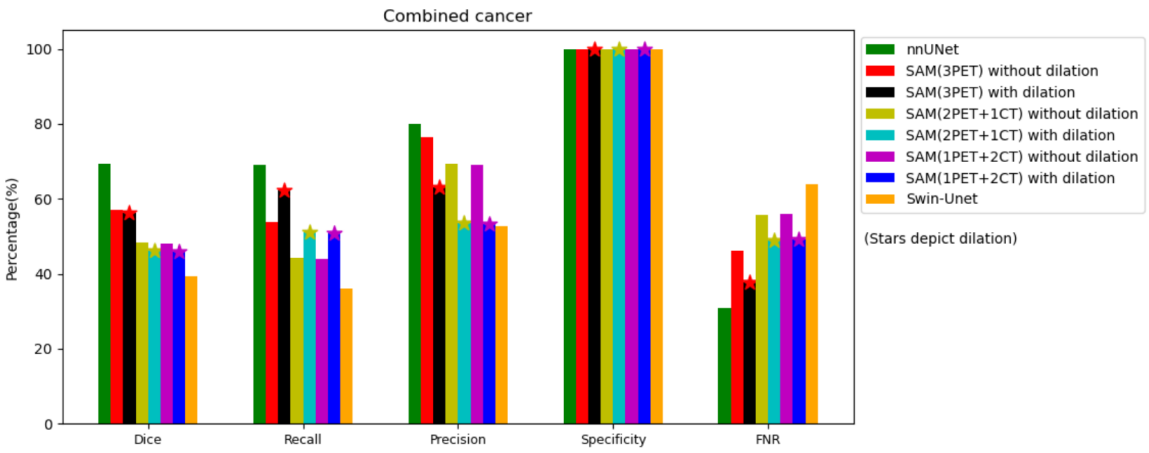

Figure 4.4: Comparison among the nnU-Net, SAM with nnU-Net prompts (indicated as SAM in the legend), and Swin-Unet on three cancers combined using five performance measures. 'Dice' denotes the Dice score

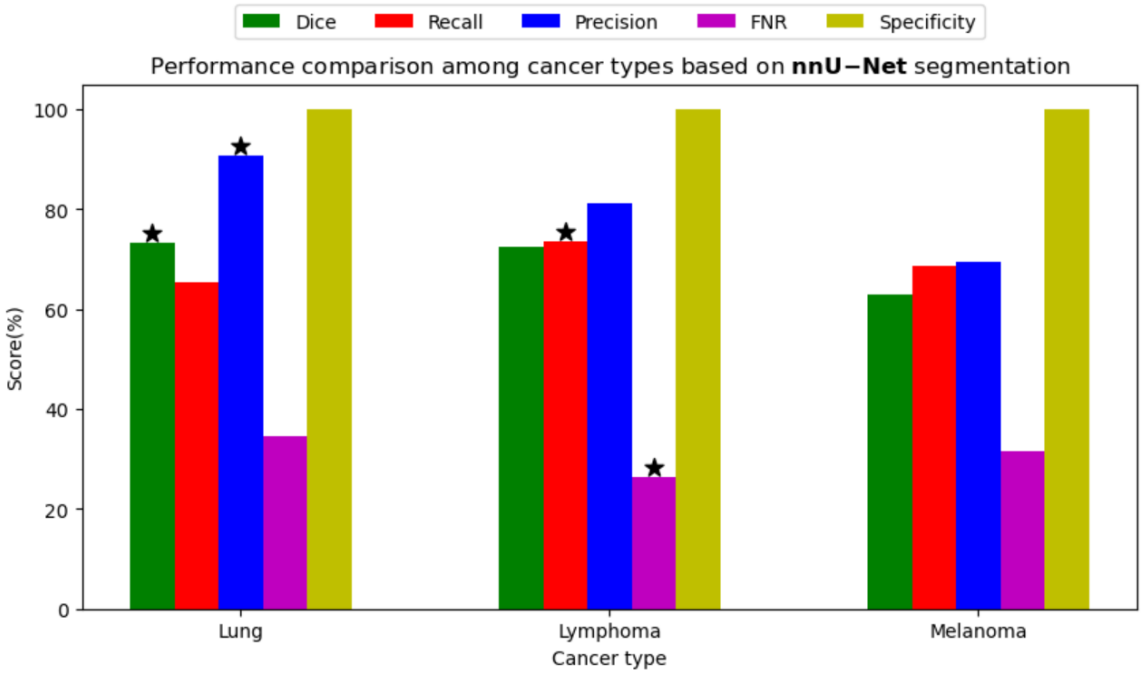

Figure 4.5: Comparison of cancer evaluation metrics for the nnU-Net method. 'Dice' denotes the Dice score, and stars indicate the best scores.

In Figure 4.5, we observe that lung cancer achieves the highest Dice and Precision scores. This indicates a more reliable tumor segmentation map for lung cancer compared to the other two cancer types (corresponding figures for other approaches are included in Appendix A). For lymphoma, we see slightly better Recall and FNR scores, suggesting a lower likelihood of missing tumor regions relative to the other cancers. However, this improved detection rate for lymphoma is observed only at the pixel level within image volumes. At the volume level, as shown in Table 4.1, lymphoma exhibits the highest miss rate (identified as fully normal) among the three cancer types, with a rate of $\frac{4}{145} \approx 2.76\%$.

Table 4.1: Number of missed image volumes by cancer type. 'SAM' refers to all the six "SAM with nnU-Net prompts" approaches.

Cancer / **Model**	**lung cancer (168)**	**lymphoma (145)**	**melanoma (188)**
nnU-Net	2	4	4
SAM	2	4	4
Swin-Unet	**0**	**0**	**0**

In nnU-Net and "SAM with nnU-Net prompts" experiments, nearly $\frac{201}{513} = 40\%$ of normal volumes were successfully flagged as normal, whereas in Swin-Unet approach, only $\frac{7}{513} = 1.5\%$ of normal volumes were successfully identified (Table 4.2). This data indicates that the Swin-Unet algorithm is inclined towards flagging volumes as positive cases, which can be helpful for not missing the normal volumes, but could be misleading in some cases, too.

Table 4.2: Number of normal image volumes successfully flagged as normal

	Number of image volumes (513 total)
nnU-Net	**201**
SAM	**201**
Swin-Unet	7

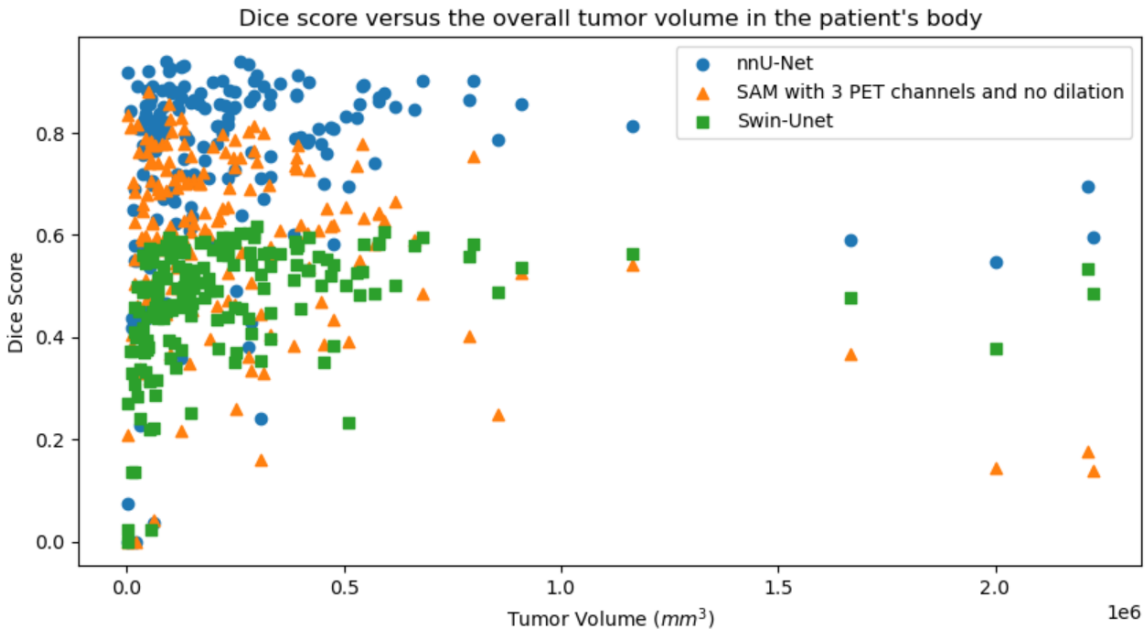

Figure 4.6: Dice score versus overall tumor volume

Figure 4.6 suggests that the performance of the three methodologies is not dependent on the total tumor spread within the body. As can be seen, the models are performing even better for volumes which have relatively smaller tumor spread. For volumes with a combined tumor spread exceeding 1000 cm^3, the three algorithms tend to show reduced performance compared to those with a total tumor spread below 1000 cm^3.

Another takeaway from this figure is that "SAM with nnU-Net prompts" is performing worse than Swin-Unet on volumes whose total tumor spread exceeds 1000 cm^3. But in low tumor spread regime, we clearly observe the superiority of nnU-Net, followed by "SAM with nnU-Net prompts", and then Swin-Unet.

4.1.2 Visual examples of segmentation

In Figures 4.7, 4.8 and 4.9, we compare the performance of three methodologies on a specific slice—the one with the largest tumor spread across all slices in the same volume. These image volumes contain the largest tumor spread compared to other image volumes of the same cancer type. We observe that the "SAM with nnU-Net prompts" algorithm, using 3 PET channels and no dilation, fails to cover most areas. In contrast, Swin-Unet demonstrates better performance, although this improvement is only evident in cases where the total tumor spread in the body exceeds cm^3.

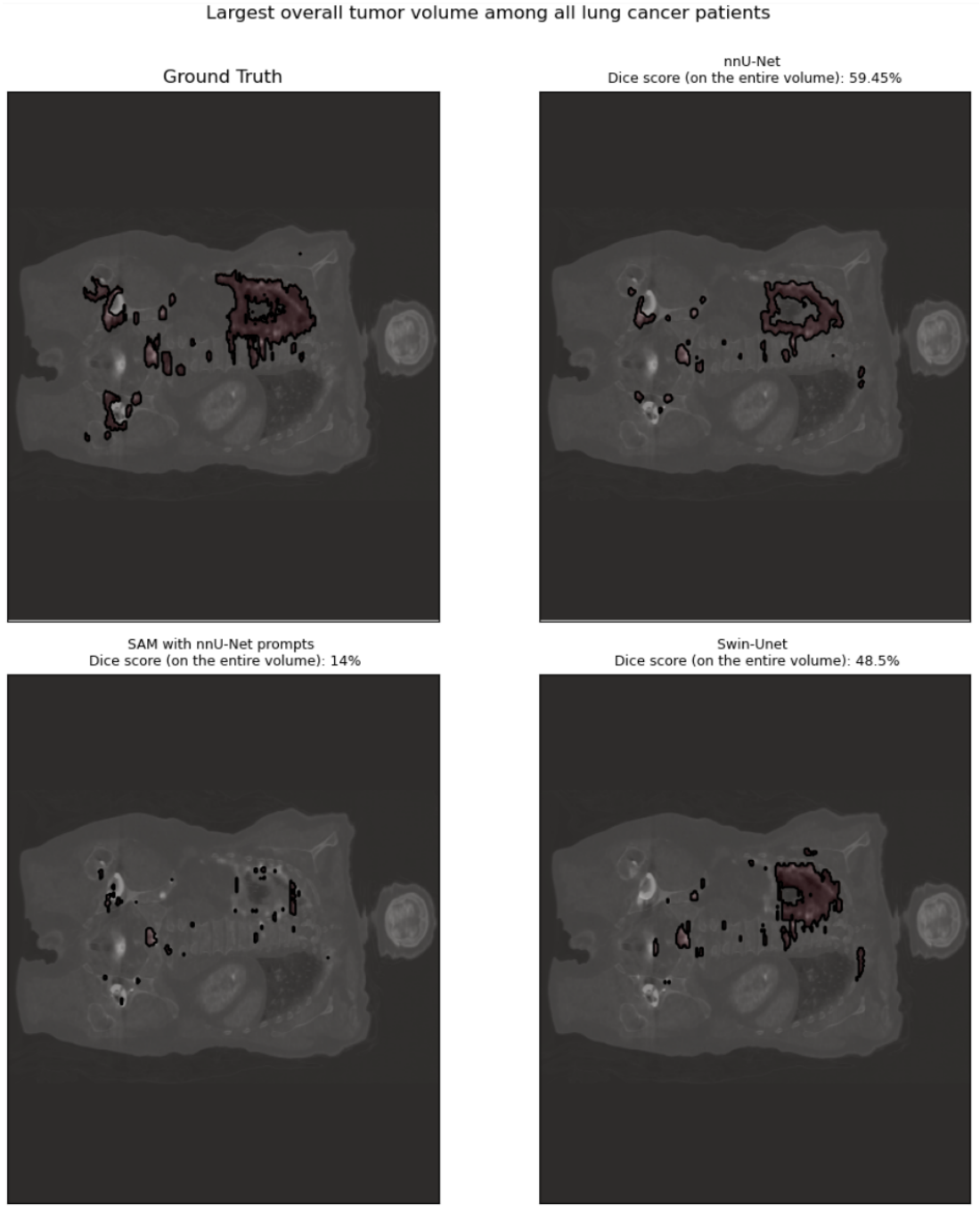

Figure 4.7: Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all lung cancer patients. This slice has the largest tumor volume of all slices for this patient.
The SAM experiment uses 3 PET channels and no dilation

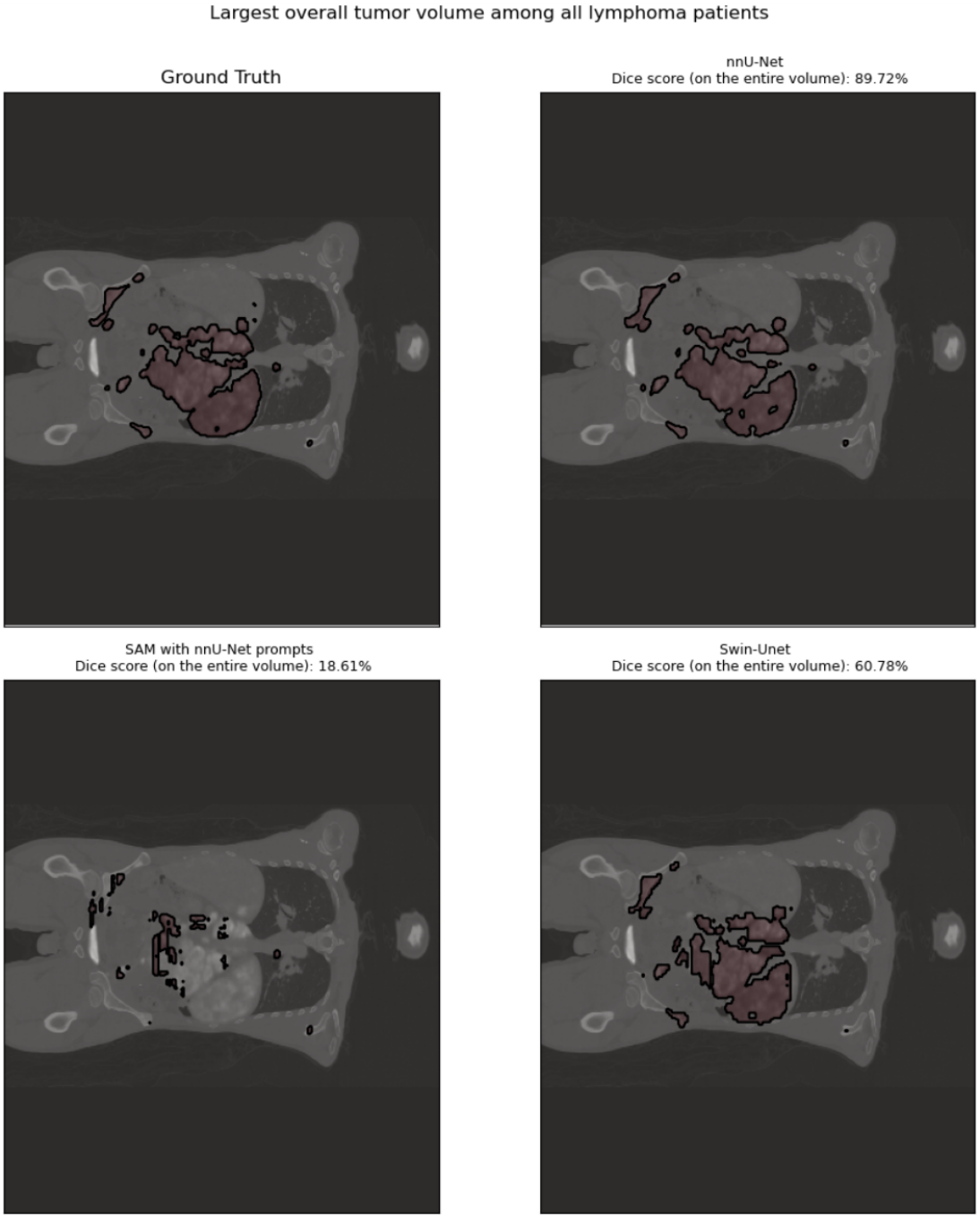

Figure 4.8: Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all lymphoma patients. This slice has the largest tumor volume of all slices for this patient.
The SAM experiment uses 3 PET channels and no dilation

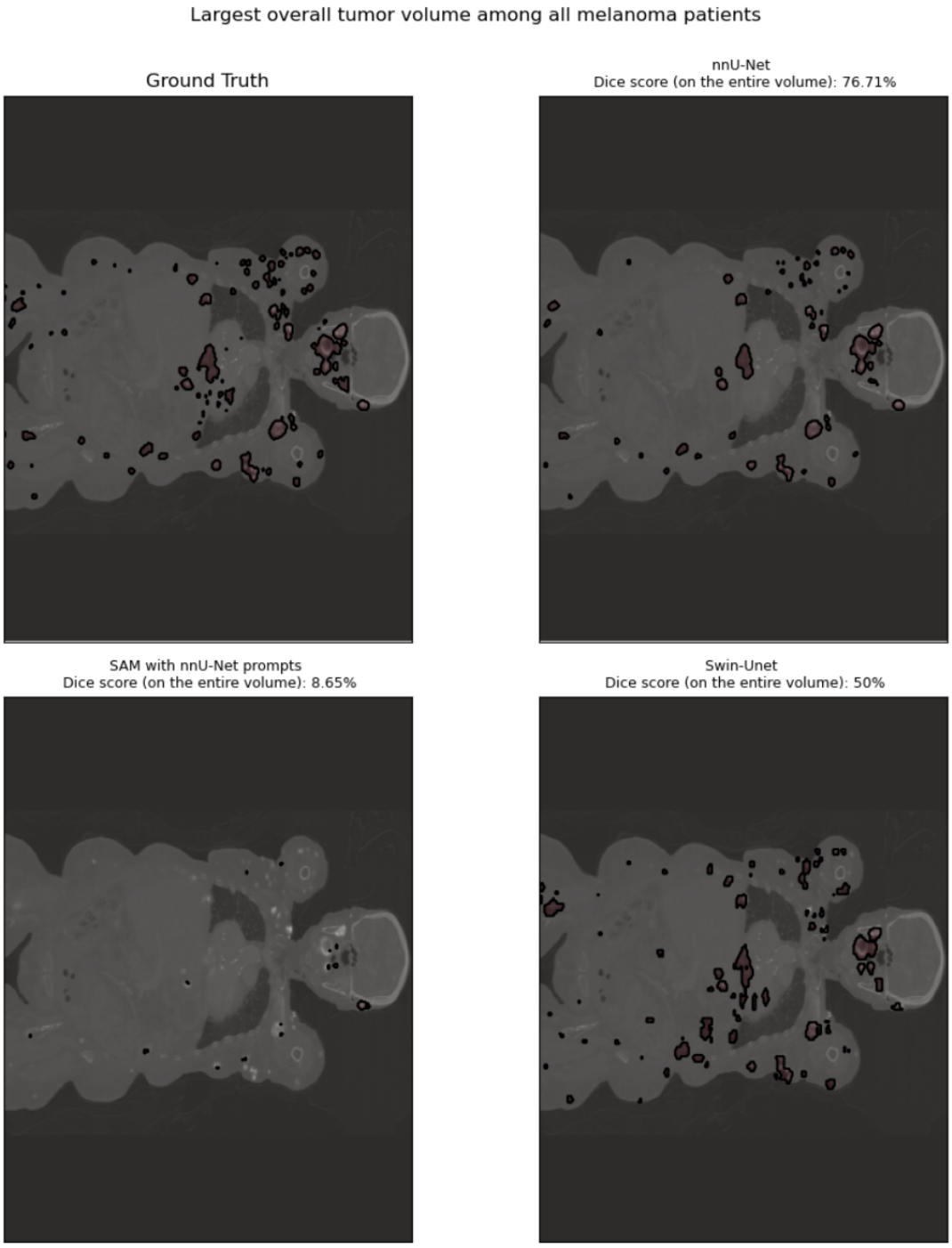

Figure 4.9: Visual comparison of the three methods on a particular slice taken from the patient with the largest tumor volume among all melanoma patients. This slice has the largest tumor volume of all slices for this patient.
The SAM experiment uses 3 PET channels and no dilation

4.2 Lesion component-wise analysis

Figure 4.10 shows the Spearman correlation coefficient measured on the number of connected components in the ground truth and the number of connected components in the prediction maps. As can be seen, for nnU-Net and "SAM with nnU-Net prompts" experiments, we observe a better correlation between the number of lesion components generated in the prediction map and that of ground truth compared to Swin-Unet. This means that Swin-Unet usually fails to follow the true number of lesion components in the image volume, which was foreseeable as it is a 2D method, compared to the other two 3D methods.

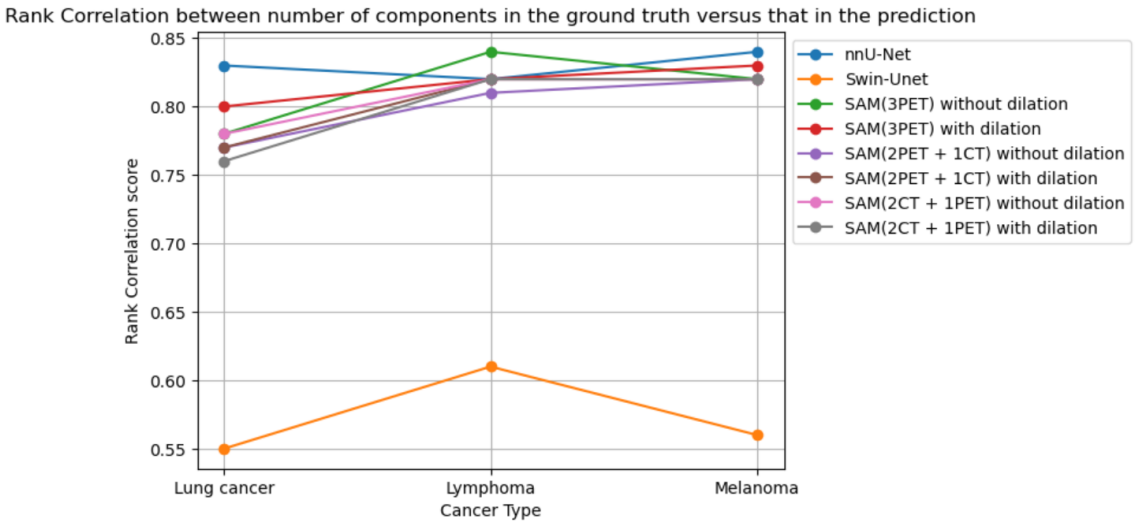

Figure 4.10: Spearman correlation between the number of connected components in the ground truth and the number of connected components in the prediction map

As shown in Tables (4.3, 4.4, 4.5, 4.6), nnU-Net consistently outperforms other algorithms in terms of median and average/highest lesion component-wise Dice scores, median volume-wise Dice scores, and the percentage of image volumes for which the

highest lesion component-wise Dice score is obtained from the largest lesion component in the body. Swin-Unet manifested comparable results to nnU-Net when delineating the largest tumor lesion in the body according to Table 4.6. SAM, however, did not have a satisfactory outcome for segmenting largest tumor lesion despite being prompted.

We also note that the volume-wise segmentation performance across cancer types and all cancers combined (Tables 4.3, 4.4, and 4.5) is closer to the median highest Dice scores rather than the median average Dice scores across the components for lung cancer and lymphoma. For all models, the largest lesion component contributed to the best segmentation in 71% to 91% of the image volumes for lung cancer. However, this was reduced for lymphoma (58% - 88%) and melanoma (57% - 76%) as shown in Table 4.6 (the accumulated number of lesion component pairs considered for Dice scores calculation is shown in Table 4.7). However, this trend only holds for 6 out of 8 models (with the exception of SAM with nnU-Net prompts with 2 PET 1 CT, and 1 PET 2 CT both with no dilation experiments). Therefore, we can state that in most cancerous image volumes, the largest component in the ground truth contributed to the highest Dice score. Given the 91% inclusion of largest components in highest lesion component-wise Dice score for nnU-Net, we can say that lung cancer biomarkers could be extracted more reliably from the largest lesion component automatically using nnU-Net using both modalities (PET/CT).

Table 4.3: Median average lesion component-wise Dice score (Range).
Bold numbers indicate the highest values obtained for each cancer type and combined.
D: dilation, ND: no dilation

Model \ Cancer type	Lung cancer	Lymphoma	Melanoma	Cancers combined
nnU-Net	**0.43** (0.07 - **0.94**)	**0.39** (0.03 - **0.95**)	**0.49** (0.02 - **0.96**)	**0.44** (0.02 - **0.96**)
SAM(3PET+ND)	0.37 (0.05 - 0.90)	0.26 (0.02 - 0.94)	0.45 (0.02 - 0.93)	0.39 (0.02 - 0.94)
SAM(3PET+D)	0.38 (0.09 - 0.88)	0.28 (0.02 - 0.90)	0.38 (0.02 - 0.91)	0.36 (0.02 - 0.91)
SAM(2PET1CT+ND)	0.26 (0.03 - 0.88)	0.19 (0.01 - 0.87)	0.40 (0.02 - 0.92)	0.31 (0.01 - 0.92)
SAM(2PET1CT+D)	0.29 (0.06 - 0.87)	0.18 (0.02 - 0.89)	0.34 (0.02 - 0.89)	0.29 (0.02 - 0.89)
SAM(1PET2CT+ND)	0.25 (0.03 - 0.89)	0.20 (0.01 - 0.89)	0.40 (0.02 - 0.90)	0.31 (0.01 - 0.90)
SAM(1PET2CT+D)	0.28 (0.03 - 0.85)	0.18 (0.01 - 0.87)	0.34 (0.02 - 0.90)	0.28 (0.01 - 0.90)
Swin-Unet	0.40 (0.15 - 0.91)	0.32 (0.03 - 0.90)	0.39 (0.02 - 0.93)	0.37 (0.02 - 0.93)

Table 4.4: Median highest lesion component-wise Dice score (Range).
Bold numbers indicate the highest values obtained for each cancer type and combined.
D: dilation, ND: no dilation

Model \ Cancer type	Lung cancer	Lymphoma	Melanoma	Cancers combined
nnU-Net	**0.86** (0.07 - **0.96**)	**0.83** (0.06 - **0.98**)	**0.85** (0.06 - **0.98**)	**0.85** (0.06 - **0.98**)
SAM(3PET+ND)	0.69 (0.21 - 0.90)	0.60 (0.09 - 0.94)	0.73 (0.42 - 0.95)	0.67 (0.09 - 0.95)
SAM(3PET+D)	0.70 (0.32 - 0.88)	0.62 (0.10 - 0.94)	0.63 (0.21 - 0.91)	0.67 (0.10 - 0.94)
SAM(2PET1CT+ND)	0.57 (0.08 - 0.88)	0.52 (0.09 - 0.92)	0.64 (0.31 - 0.92)	0.58 (0.08 - 0.92)
SAM(2PET1CT+D)	0.56 (0.09 - 0.90)	0.51 (0.10 - 0.92)	0.56 (0.12 - 0.92)	0.55 (0.09 - 0.92)
SAM(1PET2CT+ND)	0.55 (0.08 - 0.89)	0.52 (0.08 - 0.92)	0.63 (0.37 - 0.92)	0.57 (0.08 - 0.92)
SAM(1PET2CT+D)	0.55 (0.15 - 0.86)	0.51 (0.12 - 0.90)	0.57 (0.28 - 0.92)	0.54 (0.12 - 0.92)
Swin-Unet	0.78 (0.25 - 0.91)	0.70 (0.03 - 0.92)	0.62 (0.12 - 0.93)	0.72 (0.03 - 0.93)

Table 4.5: Median volume-wise Dice score (Range).
Bold numbers indicate the highest values obtained for each cancer type and combined.
D: dilation, ND: no dilation

Model \ Cancer type	Lung cancer	Lymphoma	Melanoma	Cancers combined
nnU-Net	**0.80** (0 - **0.94**)	**0.81** (0 - **0.97**)	**0.75** (0 - **0.96**)	**0.79** (0 - **0.97**)
SAM(3PET+ND)	0.62 (0 - 0.88)	0.61 (0 - 0.92)	0.64 (0 - 0.93)	0.63 (0 - 0.93)
SAM(3PET+D)	0.65 (0 - 0.85)	0.65 (0 - 0.92)	0.56 (0 - 0.90)	0.63 (0 - 0.92)
SAM(2PET1CT+ND)	0.52 (0 - 0.88)	0.48 (0 - 0.87)	0.55 (0 - 0.92)	0.51 (0 - 0.92)
SAM(2PET1CT+D)	0.53 (0 - 0.87)	0.48 (0 - 0.86)	0.44 (0 - 0.92)	0.50 (0 - 0.92)
SAM(1PET2CT+ND)	0.50 (0 - 0.89)	0.48 (0 - 0.86)	0.54 (0 - 0.92)	0.51 (0 - 0.92)
SAM(1PET2CT+D)	0.51 (0 - 0.84)	0.48 (0 - 0.85)	0.44 (0 - 0.92)	0.48 (0 - 0.92)
Swin-Unet	0.50 (0 - 0.62)	0.48 (0 - 0.61)	0.35 (0 - 0.61)	0.45 (0 - 0.62)

Table 4.6: Percentage of image volumes for which the highest lesion component-wise Dice score was obtained from the largest lesion component in the ground truth. Bold numbers indicate the highest values obtained for each cancer type and combined.
D: dilation, ND: no dilation

Cancer type / **Model**	**Lung cancer**	**Lymphoma**	**Melanoma**	**Cancers combined**
nnU-Net	**90.48**	**87.59**	**75.53**	**84.03**
SAM(3PET+ND)	80.36	68.27	62.77	70.26
SAM(3PET+D)	83.93	73.79	67.55	74.85
SAM(2PET1CT+ND)	71.43	58.62	60.11	63.47
SAM(2PET1CT+D)	71.43	62.76	59.04	64.27
SAM(1PET2CT+ND)	71.43	57.93	60.64	63.47
SAM(1PET2CT+D)	70.83	64.83	56.91	63.87
Swin-Unet	87.50	78.62	57.45	73.65

Table 4.7: Number of lesion-component pairs (lesion-component from ground truth versus that from predicted segmentation map) for which the calculated Dice scores are not zero.
D: dilation, ND: no dilation

Cancer type / **Model**	**Lung cancer**	**Lymphoma**	**Melanoma**	**Cancers combined**
nnU-Net	1488	2255	2483	6226
SAM(3PET+ND)	1667	2875	2623	7165
SAM(3PET+D)	1608	2763	2617	6988
SAM(2PET1CT+ND)	1510	2145	1783	5438
SAM(2PET1CT+D)	1418	2065	1759	5942
SAM(1PET2CT+ND)	1513	2174	1804	5491
SAM(1PET2CT+D)	1433	2081	1793	5307
Swin-Unet	1452	2581	2737	6770

Chapter 5

Discussion

In this thesis, we evaluated the performance of three deep learning-based models for segmenting tumors in a whole-body PET/CT dataset from The Cancer Imaging Archive (TCIA). The models assessed include a convolutional neural network (nnU-Net), a transformer-based model (Swin-Unet), and a hybrid model that combines both of them (SAM with nnU-Net prompts).

Our evaluations were conducted in two categories: volume-wise and lesion component-wise. The volume-wise evaluation involves computing metrics across the entire 3D volume, while the lesion component-wise evaluation focuses on calculating metrics for individual lesion components within the volume. Additionally, we compared the performance of these models across different cancer types provided in the dataset, namely lung cancer, melanoma, and lymphoma.

Overall, our results verify the superiority of nnU-Net model compared to the others. The average volume-wise Dice score over all cancer types for nnU-Net, "SAM with nnU-Net prompts with 3 PET channels and no dilation", and for Swin-Unet, are 69.22%, 57.20%, and 39.36% respectively. The rest of the "SAM with nnU-Net

prompts" experiments showed inferior results with 2 CT, 1 PET channels and dilation method maintaining the lowest Dice score of 46%. In terms of lesion component-wise analysis, we computed the Dice scores between all the isolated lesion component pairs in the ground truth and prediction maps. nnU-Net obtains a median largest lesion component-wise Dice score of 85%, followed by 67% for "SAM with nnU-Net prompts with 3 PET channels and no dilation", and 72% for Swin-Unet on three cancer types combined. The least median largest lesion component-wise Dice score belonged to "SAM with nnU-Net prompts with 2 CT, 1 PET channels and with dilation" with a Dice score of 54%.

Considering volume-wise analyses with respect to cancer types, lung cancer consistently holds the highest Precision score. Out of 8 models, 7 reported the highest Dice score on lung cancer image volumes. Out of 8 models, 6 reported the best Recall and FNR scores for melanoma image volumes, while the other two models held this title in favor of lymphoma patients. Therefore, with melanoma and lymphoma volumes, it is less likely that we miss the tumor areas, but at the cost of reduced Precision.

We compared our results with A Alloula et. al,[107] which used nnU-Net on the same dataset in AutoPET 2023 challenge[108]. They used 90% of the dataset for training and cross-validated on the remaining 10%. They reported an overall Dice score of 74.3% on the validation set without providing a breakdown of the results with respect to the cancer types. Another study[90] used nnU-Net on the same dataset, performed 5-fold cross-validation, and reported 73% mean Dice score over the positive cases. Our overall mean Dice score over the positive cases, was 69.22%. The deviation among the results could possibly stem from the different data split between folds considered in our experiments and theirs.

The nnU-Net algorithm successfully captures most tumor areas, accurately identified normal patients (201 out of 513), and infrequently flagged normal patients as positive. While the model is well-documented and explained on the authors' GitHub page, some fundamental information about the model remain vague. For instance, the process of computing the pseudo Dice score is not well-explained; it is unclear how random patches are sampled and in what proportion to the size of the validation set. Additionally, it is not specified whether an internal validation set is used when performing 5-fold cross-validation. It remains uncertain whether the pseudo Dice score closely aligns with the actual Dice score, regardless of the imaging modality and data imbalance in the datasets.

The "SAM with nnU-Net prompts" experiments demonstrated a close relationship with nnU-Net but performed poorly for patients with large tumor volumes (Figure 4.6). This issue is particularly concerning for lung cancer patients, as the largest tumor components in these patients are typically located in the lungs. Random dilation did not yield the expected results; while it improved the Dice, Recall, and FNR scores, it decreased Precision. Therefore, its utilization is contingent upon specific use case. Though there is a very recent fine-tuned version of SAM for medical images[28] (called MedSAM); it is already trained on the same PET/CT dataset from TCIA, making it unsuitable for evaluation in our study.

Chapter 6

Limitations

There were several limitations associated with this study. For instance, Swin-Unet is a novel architecture that combines the skip-connections inherited from U-Net with a transformer. This model took significantly longer to train compared to nnU-Net (one month versus five days). We attempted to train a 3D version of Swin-Unet, but our hardware could not support large batch sizes due to GPU RAM limitations. To address this issue, we employed a method called "gradient accumulation"[109] where instead of updating the model's weights after every batch, we passed mini-batches and updated the model's weights after a specific number of time-steps (in this case, the number of slices in each image volume). However, the first fold of this 3D version of Swin-Unet model (trained on the same dataset as the 2D Swin-Unet in this thesis, subsection 3.5.1) did not perform adequately. Consequently, we halted this experiment and ruled it out, as training the model on the entire dataset would have taken considerably more time on our servers. We conducted stratified 5-fold cross-validation to ensure a balanced number of positive and negative slices in each fold. The 2D version of the model performed decently for patients with tumors

spread throughout their body but underperformed for low-volume tumors, where SAM showed better results. Additionally, Swin-Unet tended to flag most image volumes as PET-positive. The performance of Swin-UNet can improve with better experimental set-up, including hardware.

Lesion-wise analysis was a crucial aspect of this research, as it aids in the evaluation of quantitative imaging biomarkers—a subject less thoroughly investigated compared to other areas. We conducted several analyses, including calculating Dice scores between all connected component pairs in the ground truth and the prediction maps. For example, if the ground truth contained 5 connected components and the prediction map had 3, we performed $5 \times 3 = 15$ Dice score calculations for that specific volume. During lesion component-wise analysis, we relied heavily on the connected component analysis algorithm to identify the lesion components in the absence of expert-provided ground truth for connected lesions. If individual lesions in image volumes were labeled distinctly by experts apriori, we could have approached the segmentation problem from the perspective of instance segmentation. Within the stipulated time frame of this research work and the limited clinical data available, it was not feasible to obtain adequate information on the instances of the lesions. In the absence of expert-provided instance labels for the lesions, we mapped the ground truth lesions to the predicted lesions (from each of the eight models) based on the overlap of lesion components between the ground truth and the predicted segmentations.

We are obtaining better results using PET alone or in combination with CT (PET/CT), but not with CT alone. Since we are unaware whether metabolic prioritization was applied to the provided ground truth labels, we cannot comment on

whether this is a factor contributing to the better performance

Overall, the research works we reviewed on tumor segmentation in PET/CT images using automated approaches predominantly focused on the engineering aspects, with less emphasis on medical interpretation and collaboration with medical experts to refine their models for real-world applications. Therefore, we advocate for the availability of newer and richer datasets for developers and scientists to construct more robust and clinically meaningful models. An increased variety of datasets would facilitate the testing of algorithms on out-of-distribution data, thereby enabling a better assessment of the models' generalization capabilities.

We also recognized some potential architectures and solutions that could have been analyzed for the problem of tumor segmentation using PET/CT modalities (specially transformer-based models), namely Trans-Unet [86], UNETR [110], and Segmenter [111].

Transformer-based models, equipped with the self-attention mechanism, are theoretically more powerful than CNNs due to their ability to capture long-range relations and dependencies in the data. However, our results demonstrated the superiority of nnU-Net, which is based on CNNs. This discrepancy could be attributed to the dataset size, as transformer-based models tend to perform better with larger datasets, whereas models with inductive biases like CNNs perform better with smaller datasets. Additionally, our experimental setup may have influenced the outcomes. Studies suggest that the combination of transformers and U-Net is suitable for segmentation[112]. This could be a potential direction for future work stemming from this thesis.

Chapter 7

Conclusion and Future Works

This thesis evaluated the performance of three recent deep learning-based approaches for segmenting tumors/lesions in whole-body PET/CT images from The Cancer Imaging Archive (TCIA) dataset. The models assessed were nnU-Net (a CNN-based approach), Swin-Unet (a transformer-based model), and a hybrid model combining nnU-Net with the Segment Anything Model (SAM). Our evaluations were conducted at both the volume-wise and lesion component-wise levels, across three cancer types: lung cancer, melanoma, and lymphoma, as well as normal cases present in the dataset.

The nnU-Net algorithm demonstrated superior performance overall, achieving the highest volume-wise Dice scores and successfully identifying the majority of tumor areas. It was particularly effective at identifying normal patients by flagging fewer false positives compared to the other models. In contrast, Swin-Unet showed a tendency to over-segment, marking many tissue areas as PET-positive, and needed significantly more training time given the available hardware resources. The hybrid approach using "SAM with nnU-Net prompts" showed potential but underperformed for patients

with large tumor volumes, especially lung cancer. The nnU-Net and SAM models incorrectly flagged 2, 4, and 4 image volumes of lung cancer, lymphoma, and melanoma patients as normal, respectively, whereas Swin-Unet did not miss any malignant volume. This indicates that the performance of each model varies across cancer types, with melanoma and lymphoma patients being the most susceptible to missed detections. Out of 513 normal volumes, 201 were successfully identified by nnU-Net and SAM, but only 7 were accurately flagged as normal by Swin-Unet. Therefore, nnU-Net and "SAM with nnU-Net prompts" models show superiority in terms of correctly identifying the normal patients.

Our analyses revealed the importance of model selection based on specific application needs and highlighted the limitations of current datasets and methodologies. While nnU-Net performed well overall, the results suggest that more robust and varied datasets are needed to improve the generalizability of these models. Additionally, the unavailability of instance-related details in the ground truth segmentations limited the scope of our methodology and analysis. In the future, we expect richer datasets to emerge, where the problem of instance segmentation in different cancer types can be handled more efficiently by better models.

Appendix A

A.1 nnU-Net model evaluation on partial dataset

The details of the training procedure for this partial dataset are listed below:

- Number of patients: 270 (297 image volumes).
- Training set: 231 volumes; Testing set: 66 volumes (approximately a 78/22% split).
- The dataset split is depicted in Table A.1.
- No data augmentation was performed, as recommended by the authors.
- The original anisotropic spacing was modified from $[2.03642, 2.03642, 3]$, mm to isotropic $[2, 2, 2]$, mm spacing.
- We utilized 5-fold cross-validation to mitigate the risk of overfitting and reduce bias.
- Experiments were conducted over 1,000 epochs.

Split / Category	Train	Test
Normal images	111	25
Malignant images	120	41

Table A.1: Train/test split in partial dataset for nnU-Net

- A total of 15 models were trained for performance comparison, involving 5 loss functions and 3 variations in modalities.

At the end of training each model, two types of model weights are saved; one with the lowest training loss saved as *Best*, and one saved at the end of the 1000^{th} epoch saved as *Final*.

Loss function	**Test/Final**	**Test/Best**
Focal	11	6
Dice	4	1
CE	11	2
Tversky	2	1
Dice+CE	1	1

Table A.2: Normal cases successfully identified using PET only (out of 25)

Loss function	**Test/Final**	**Test/Best**
Focal	1	1
Dice	1	0
CE	1	1
Tversky	0	0
Dice+CE	0	0

Table A.3: Malignant cases identified as normal using PET only (out of 111)

Loss function	**Dice**		**Precision**		**Recall**	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	45.4	**41.44**	60.3	**54.27**	61.96	62.57
Dice	42.32	37.39	51.26	43.67	**67.63**	66.75
CE	**47.75**	39.97	**63.56**	48.67	62.66	63.74
Tversky	39.84	39.04	46.12	43.81	67.1	**69.25**
Dice+CE	38.29	36.78	44.65	43.53	65.71	65.68

Table A.4: Performance evaluation metrics for partial dataset trained on PET only, evaluated on the entire dataset

Loss function	Dice		Precision		Recall	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	60.9	60.65	81.4	**80.05**	61.96	62.57
Dice	64	59.29	78.18	70.96	**67.63**	66.75
CE	**64.04**	**62.39**	**85.8**	76.65	62.66	63.74
Tversky	62.19	61.89	72	69.46	67.1	**69.25**
Dice+CE	60.7	58.31	70.8	69.02	65.71	65.68

Table A.5: Performance evaluation metrics for partial dataset trained on PET only, evaluated on positive volumes only

Loss function	Test/Final	Test/Best
Focal	17	17
Dice	0	0
CE	21	21
Tversky	0	0
Dice+CE	0	0

Table A.6: Normal cases successfully identified using CT only (out of 25)

Loss function	Test/Final	Test/Best
Focal	10	11
Dice	0	0
CE	10	10
Tversky	0	0
Dice+CE	0	0

Table A.7: Malignant cases identified as normal using CT only (out of 41)

Loss function	Dice		Precision		Recall	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	27.59	25.64	57.15	59.69	27.97	25.45
Dice	15.14	15.23	15.74	15.56	**30.35**	31.01
CE	**30.86**	**29.22**	**60.72**	**63.68**	28.66	27.34
Tversky	16.23	14.7	17.24	15.27	30.02	29.72
Dice+CE	15.64	15.49	16.44	15.87	30.16	**31.81**

Table A.8: Performance evaluation metrics for partial dataset trained on CT only, evaluated on the entire dataset

Loss function	Dice		Precision		Recall	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	32.98	30.64	**71.9**	**75.6**	27.97	25.45
Dice	24.37	24.52	25.33	25.05	**30.35**	31.01
CE	**33.87**	**32.07**	68.56	71.9	28.66	27.34
Tversky	26.13	23.66	27.75	24.75	30.02	29.72
Dice+CE	25.18	24.94	26.47	25.54	30.16	**31.81**

Table A.9: Performance evaluation metrics for partial dataset trained for CT only, evaluated on positive volumes only

Loss function	Test/Final	Test/Best
Focal	12	8
Dice	0	2
CE	10	5
Tversky	7	8
Dice+CE	4	5

Table A.10: Normal cases successfully identified using PET/CT (out of 25)

Loss function	Test/Final	Test/Best
Focal	2	2
Dice	0	0
CE	1	1
Tversky	1	1
Dice+CE	0	0

Table A.11: Malignant cases identified as normal using PET/CT (out of 41)

Loss function	Dice		Precision		Recall	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	**49.5**	**45.86**	**62.79**	**55.79**	64.74	**67.49**
Dice	37.78	39.86	44.01	46.96	**66.94**	66.24
CE	47.96	44.47	62.04	54.68	64.39	67.41
Tversky	42.83	43.21	53.14	54.15	64.2	63.62
Dice+CE	40.27	38.31	47.29	43.89	66.07	64.61

Table A.12: Performance evaluation metrics for partial dataset trained on PET/CT, evaluated on the entire dataset

Loss function	Dice		Precision		Recall	
	Test/Final	Test/Best	Test/Final	Test/Best	Test/Final	Test/Best
Focal	65.2	64.88	83.72	80.1	64.74	**67.49**
Dice	60.82	62.22	70.85	73.31	**66.94**	66.24
CE	**65.51**	**66.17**	**85.3**	**82.02**	64.39	67.41
Tversky	61.63	61.13	77.06	77.17	64.2	63.62
Dice+CE	60.89	57	71.51	65.3	66.07	64.61

Table A.13: Performance evaluation metrics for partial dataset trained on PET/CT, evaluated on positive volumes only

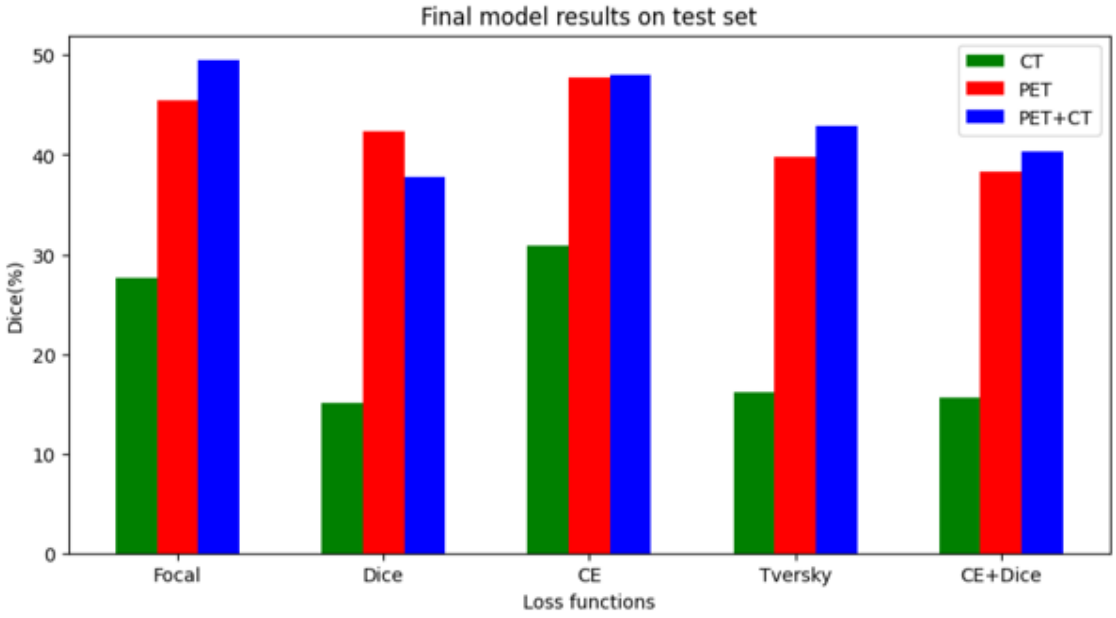

Figure A.1: Modality and loss function performance comparison for the partial dataset using nnU-Net

The tables and figure above indicate that the highest scores were achieved using the PET/CT combination with Focal loss, as shown in Table A.12. Although CE loss produced better results in some instances compared to Focal loss, the Dice score was the most critical metric, especially given that the dataset comprises both normal and malignant volumes.

A.2 Performance metrics comparison for SAM and Swin-Unet with respect to cancer type in whole-body volumes

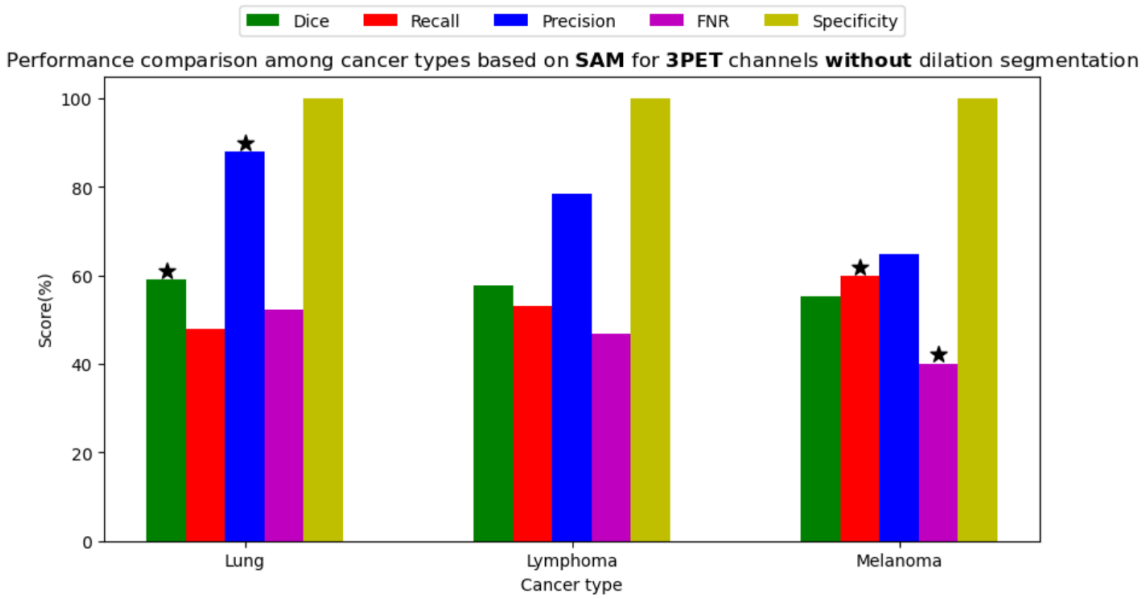

Figure A.2: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 3 PET channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score.

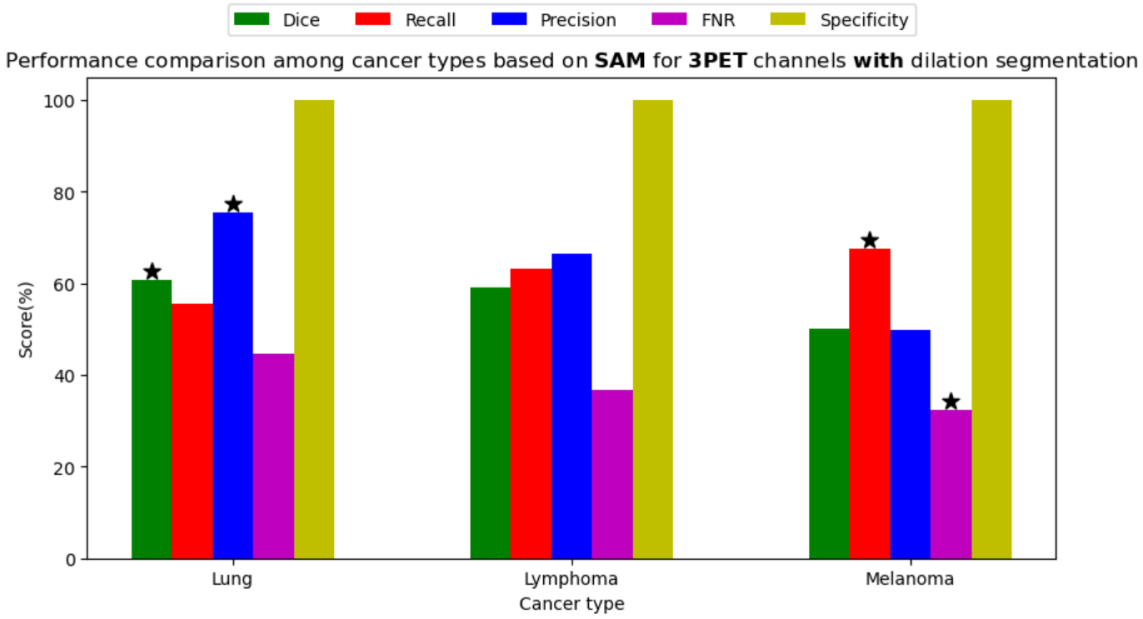

Figure A.3: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 3 PET channels and dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score.

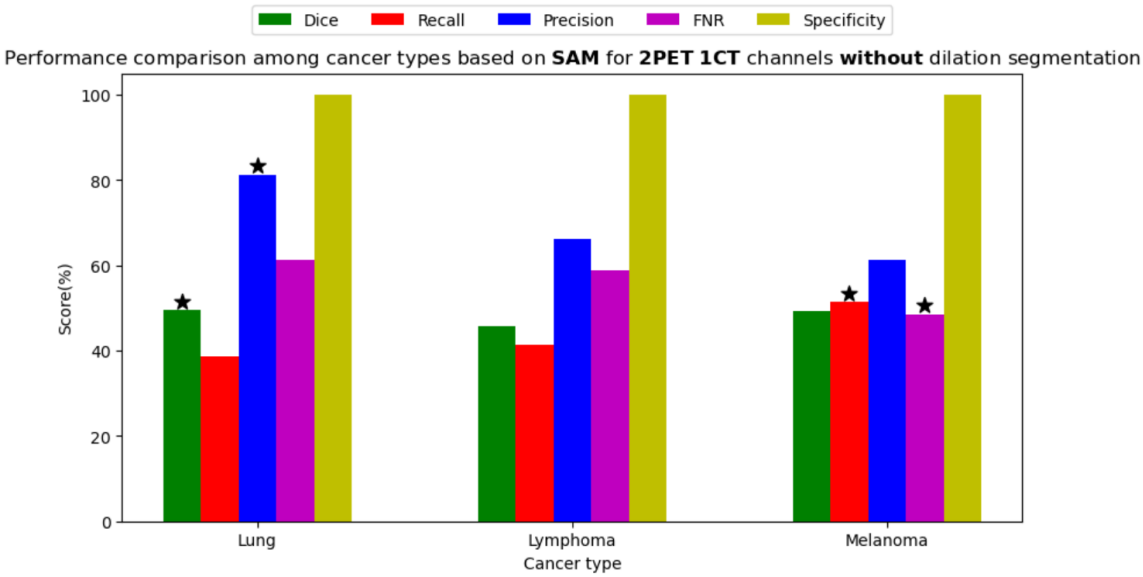

Figure A.4: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 2 PET, 1 CT channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score.

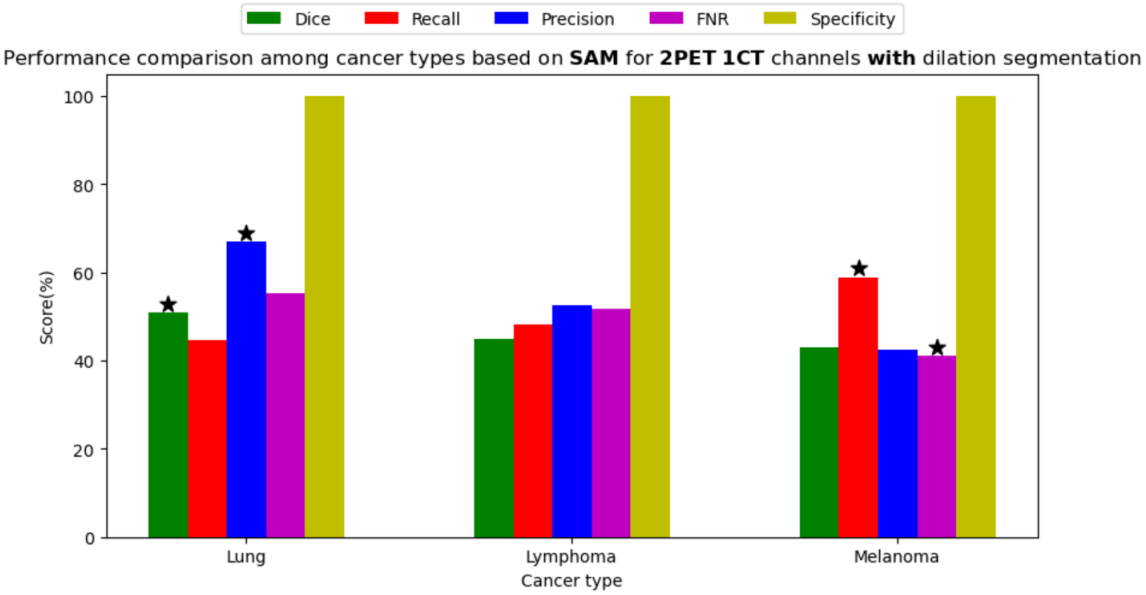

Figure A.5: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 2 PET, 1 CT channels and dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score.

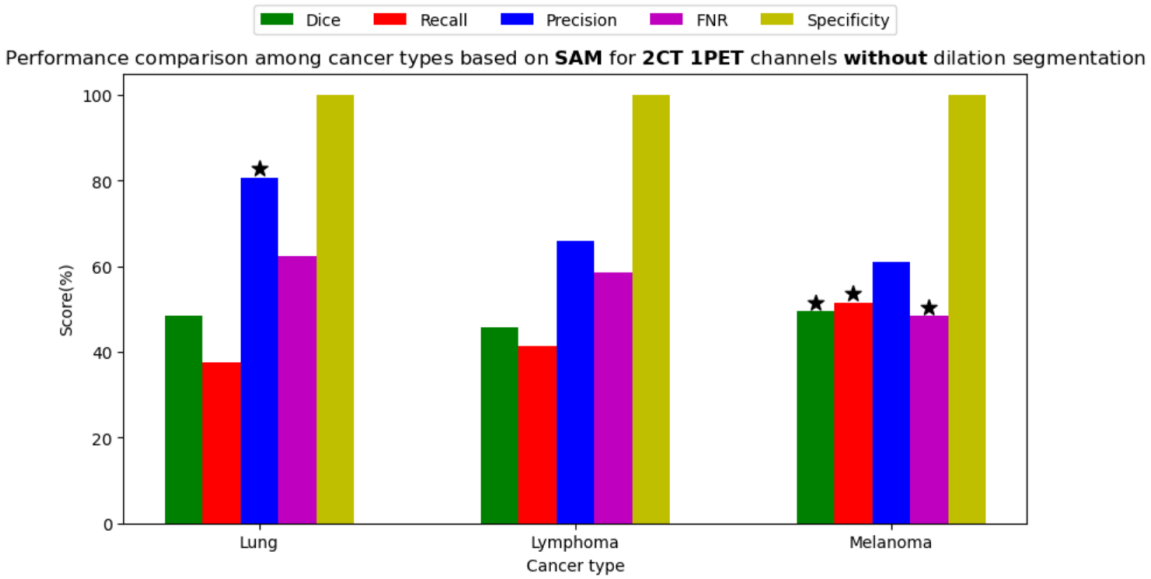

Figure A.6: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 1 PET, 2 CT channels and no dilation" method. 'Dice' denotes the Dice score, and stars indicate the highest score.

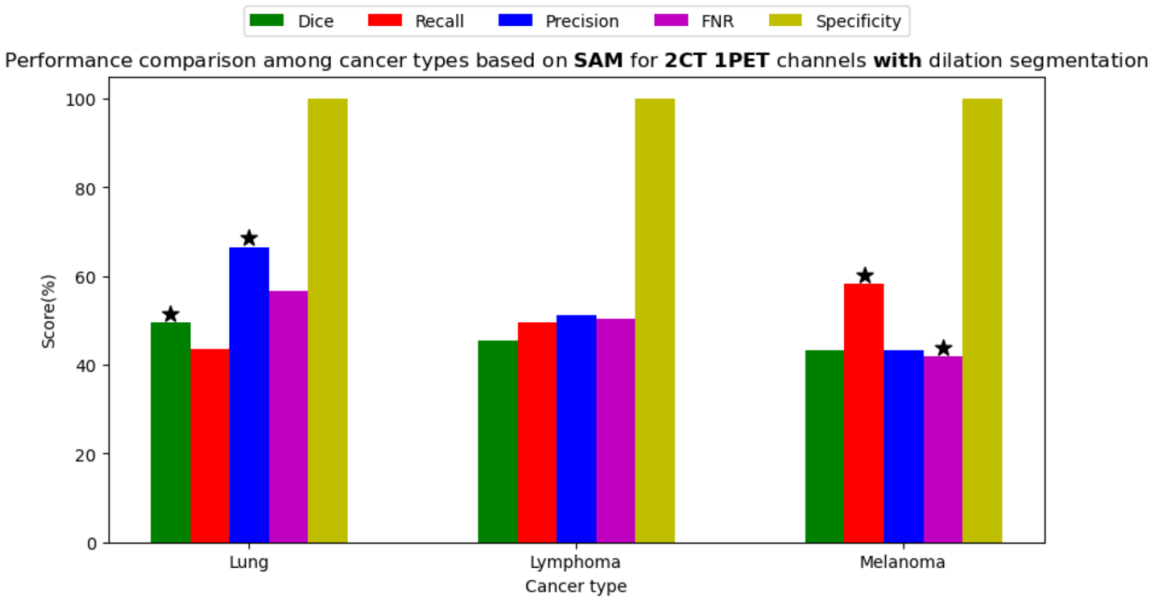

Figure A.7: Comparison of cancer evaluation metrics for the "SAM with nnU-Net prompts with 1 PET, 2 CT channels and dilation" method. 'Dice' denotes the Dice score, with and stars indicate the highest score.

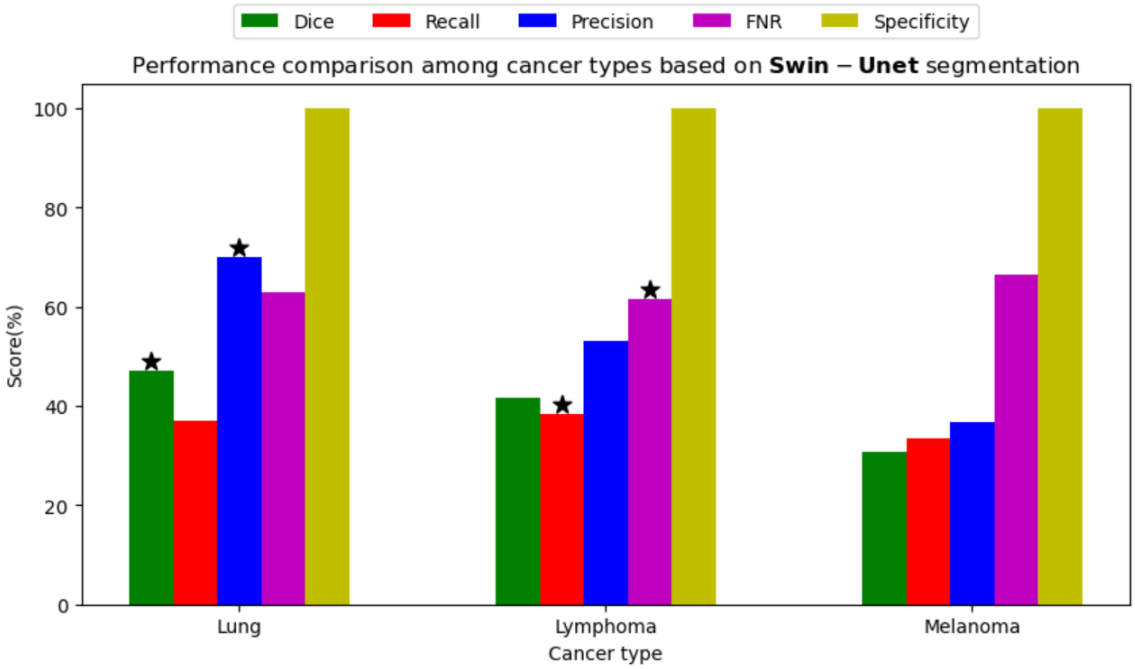

Figure A.8: Comparison of cancer evaluation metrics for the Swin-Unet method. 'Dice' denotes the Dice score, with and stars indicate the highest score.

A.3 Further insight on the dataset and nnU-Net's performance

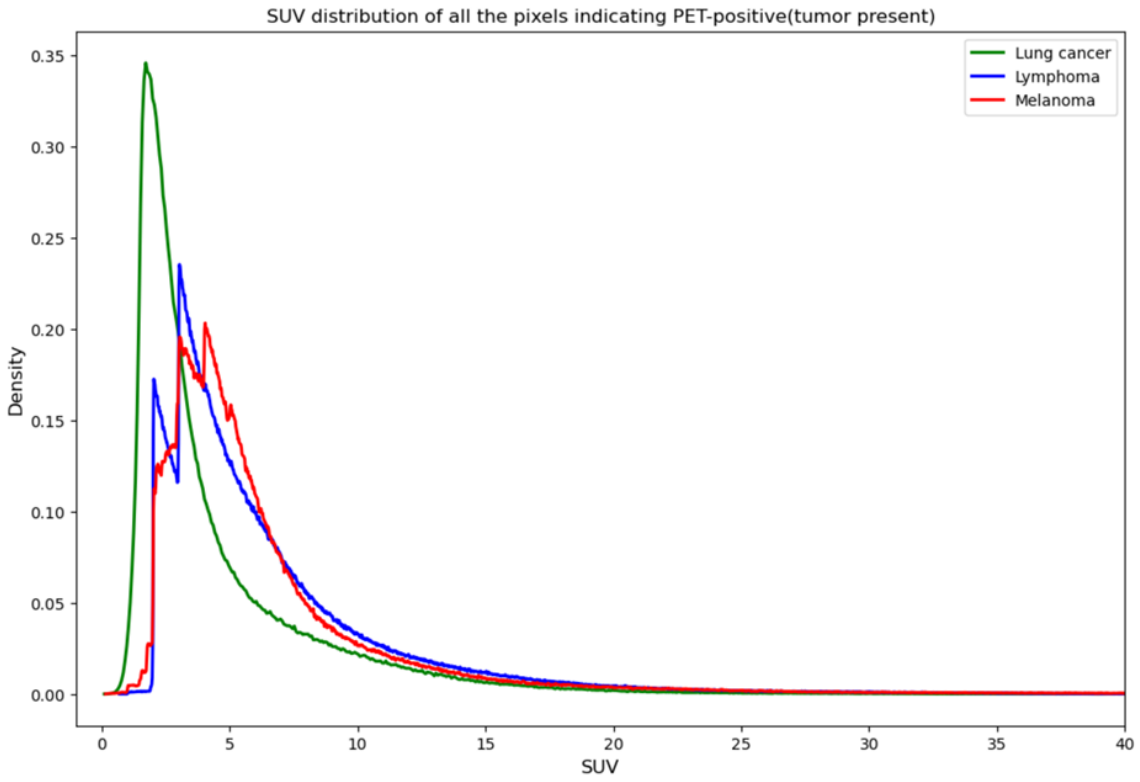

Figure A.9: Distribution of SUVs corresponding to PET-positive malignancy (presence of cancer) across three cancer types in ground truth.

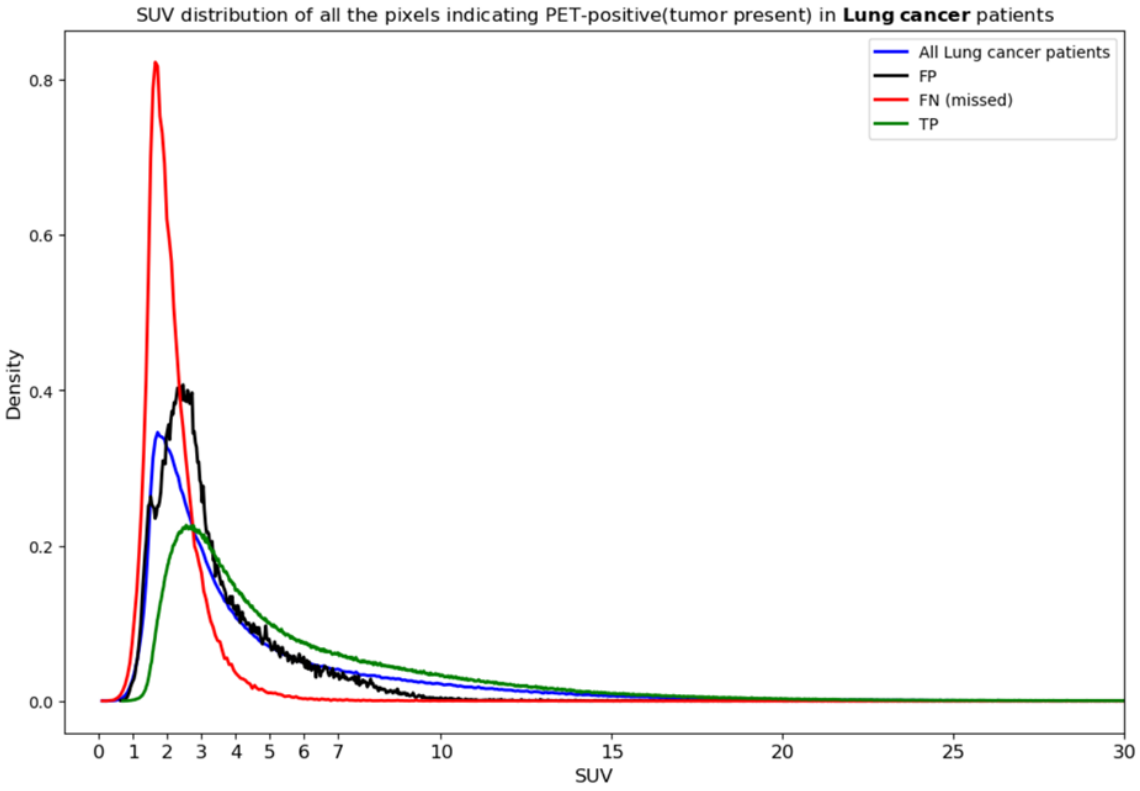

Figure A.10: Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in lung cancer patients for nnU-Net algorithm.

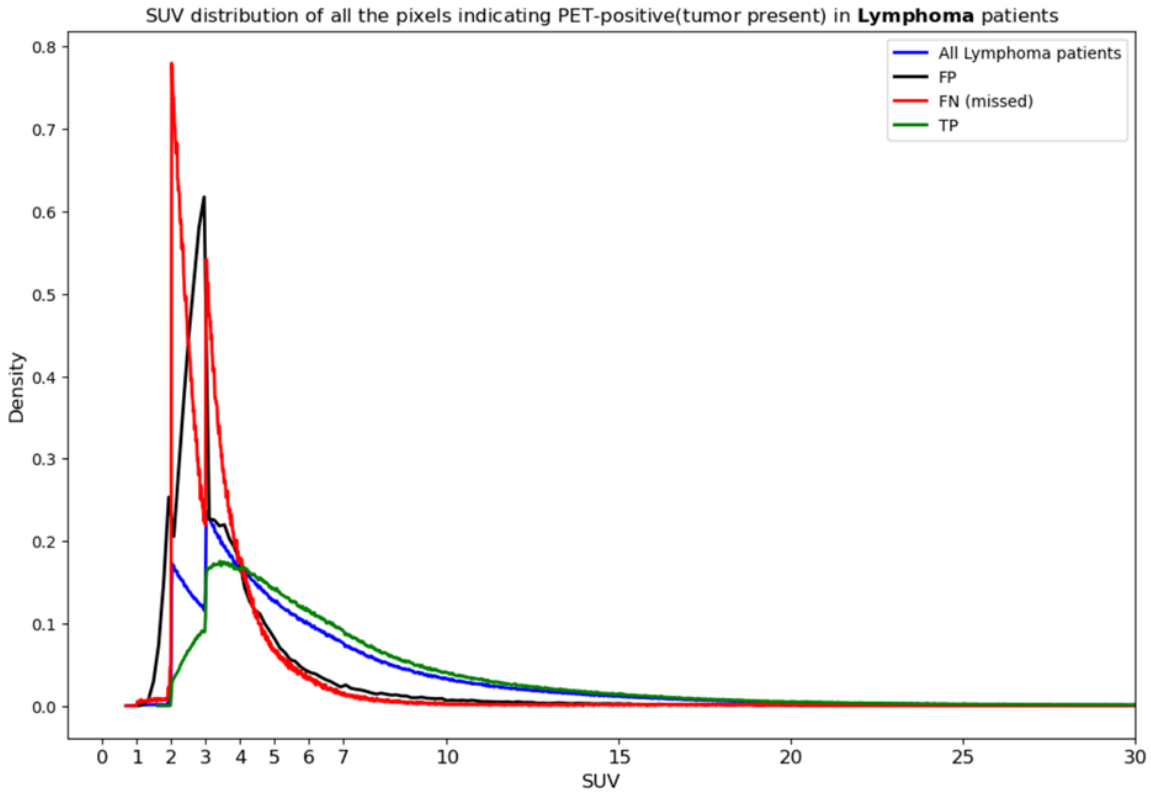

Figure A.11: Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in lymphoma patients for nnU-Net algorithm.

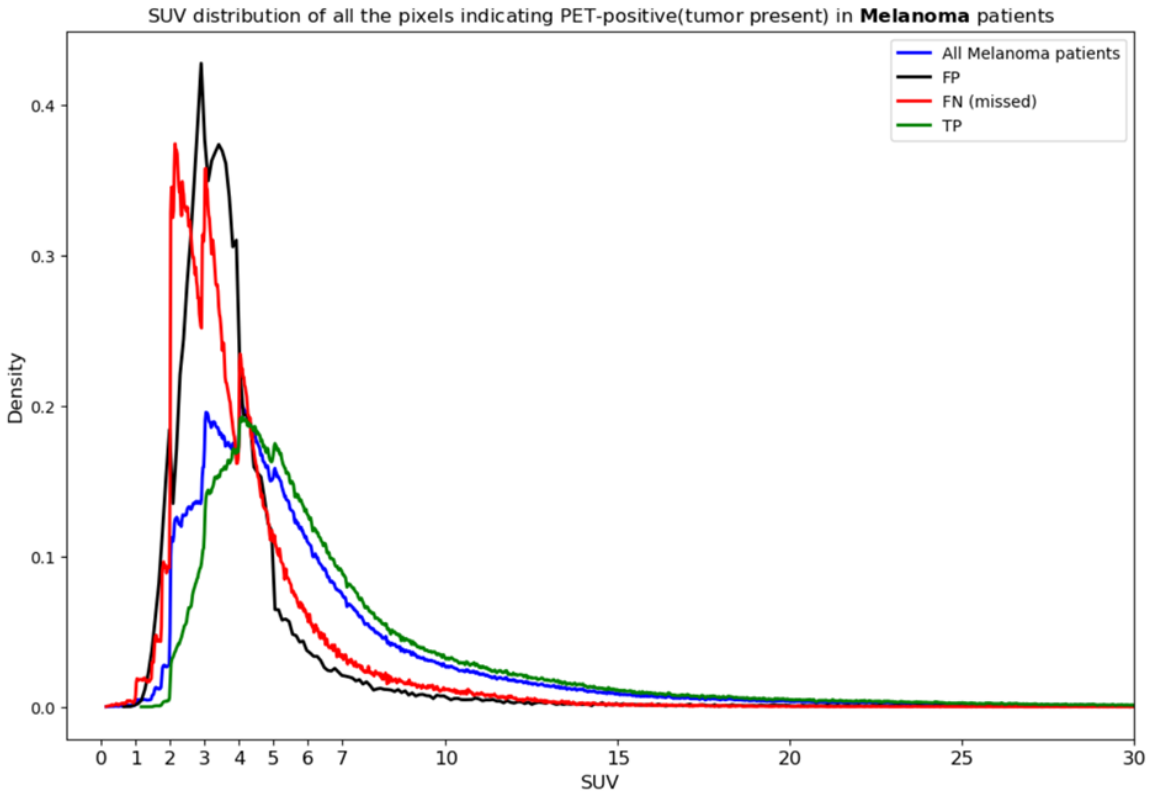

Figure A.12: Comparison of True Positive (TP), False Negative (FN) or missing, and False Positive (FP) pixels distribution in melanoma patients for nnU-Net algorithm.

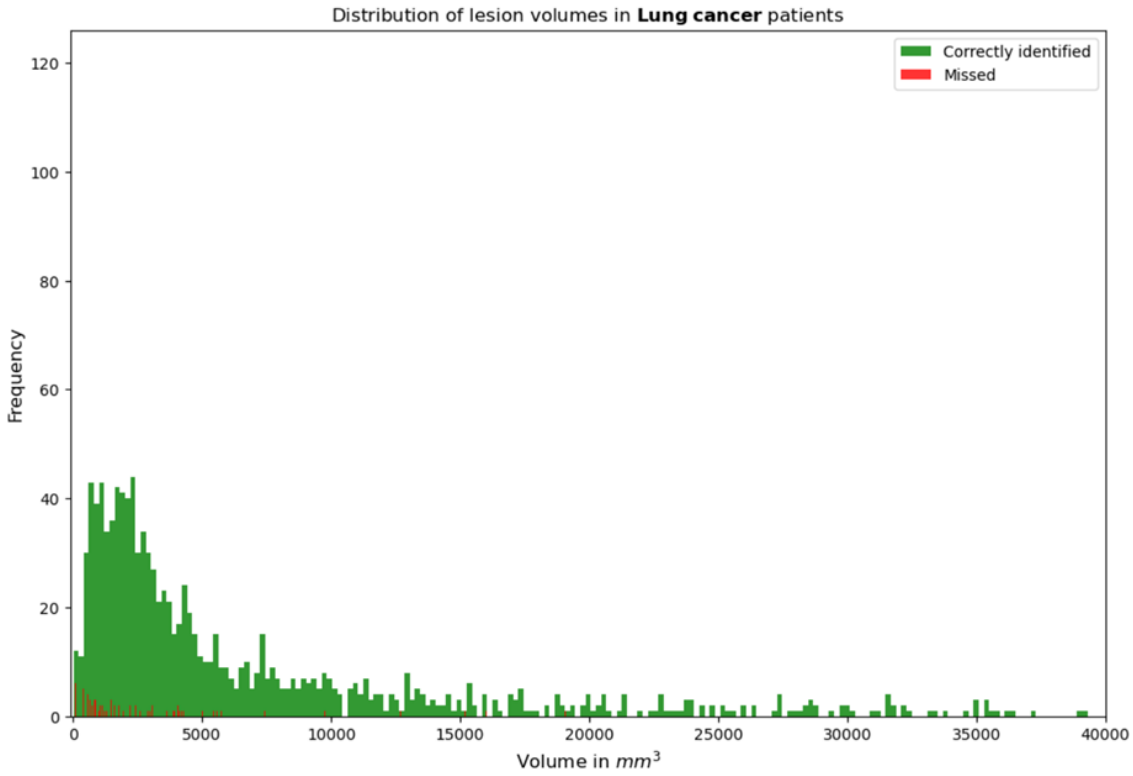

Figure A.13: Histogram of volumes of all the lesions (ground truth) in lung cancer patients for nnU-Net algorithm.

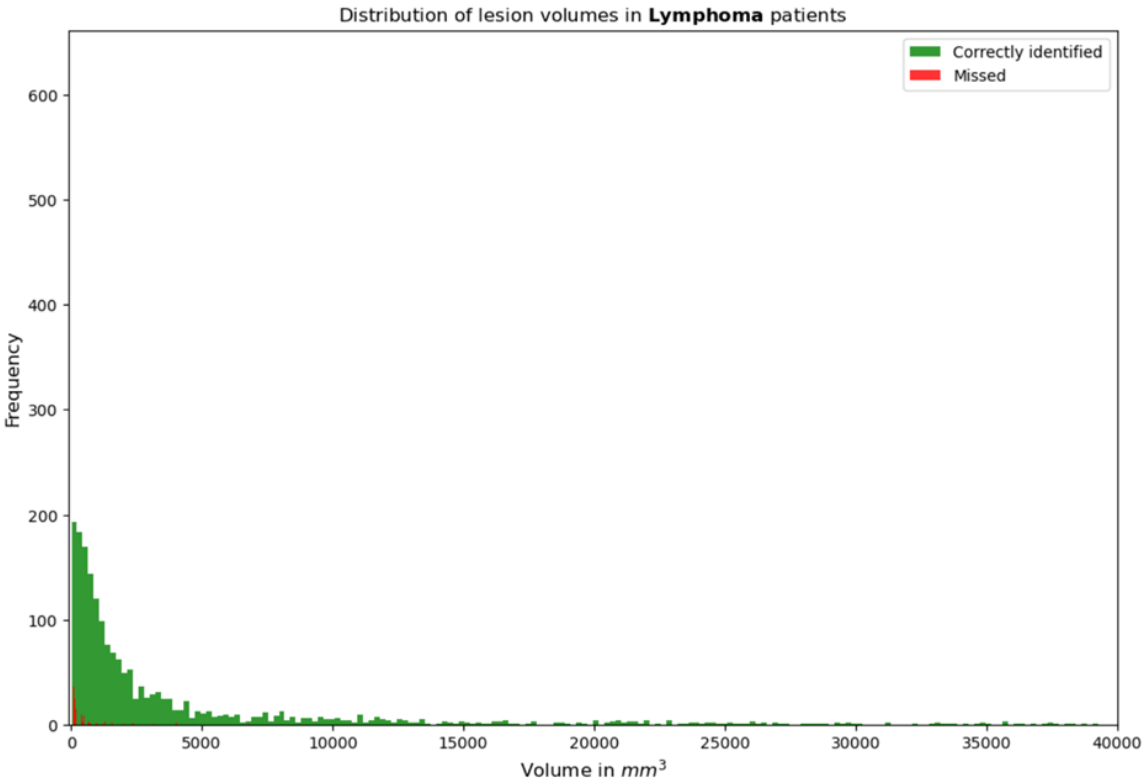

Figure A.14: Histogram of volumes of all the lesions (ground truth) in lymphoma patients for nnU-Net algorithm.

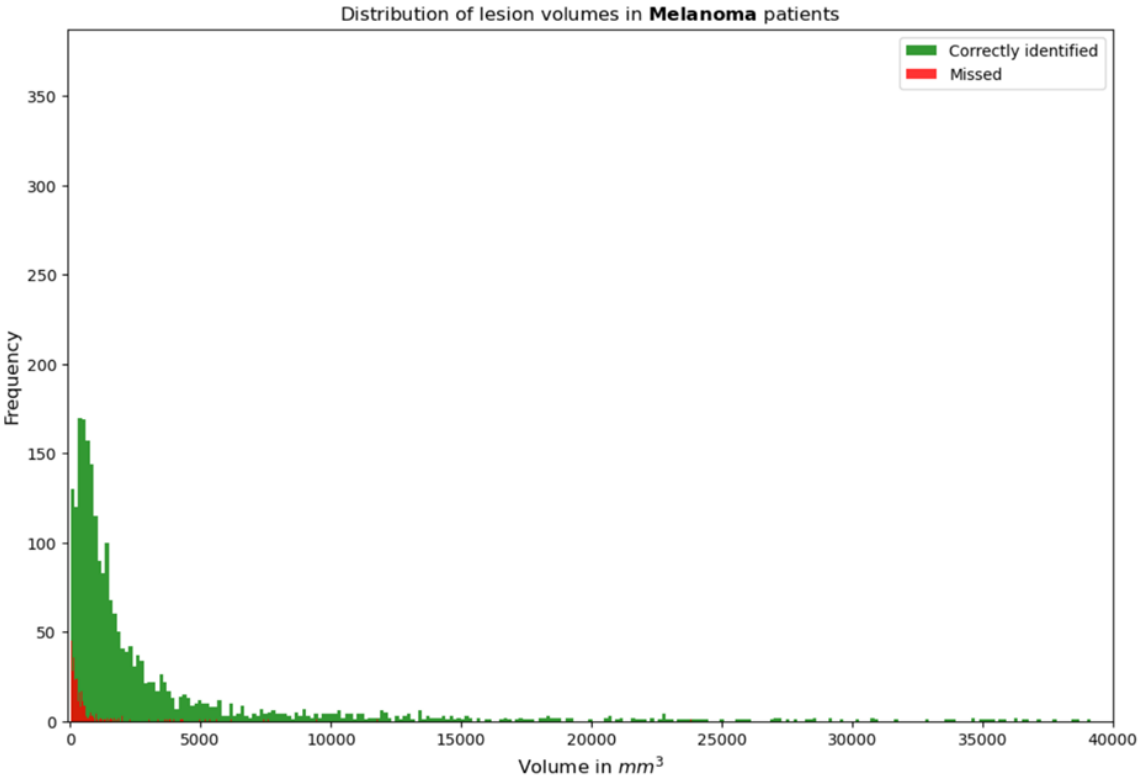

Figure A.15: Histogram of volumes of all the lesions (ground truth) in melanoma patients for nnU-Net algorithm.

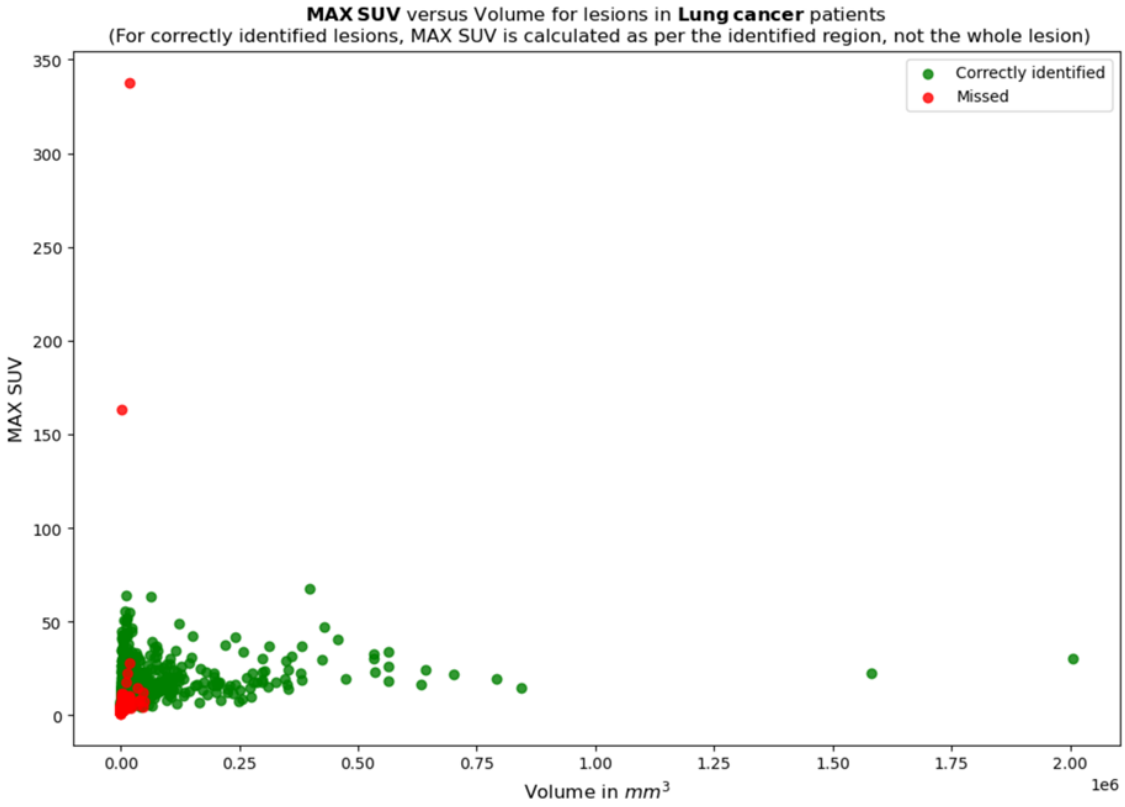

Figure A.16: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lung cancer patients.

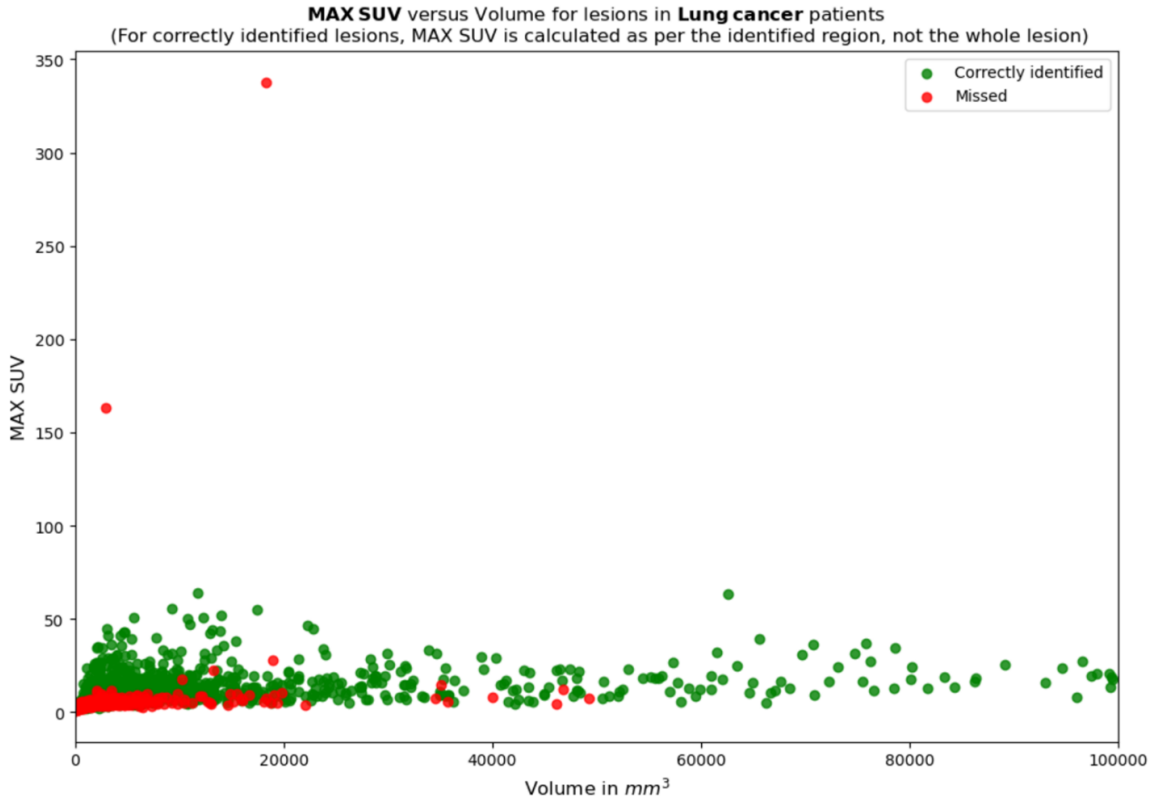

Figure A.17: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lung cancer patients with limited x-axis for a better view on the low volume lesions.

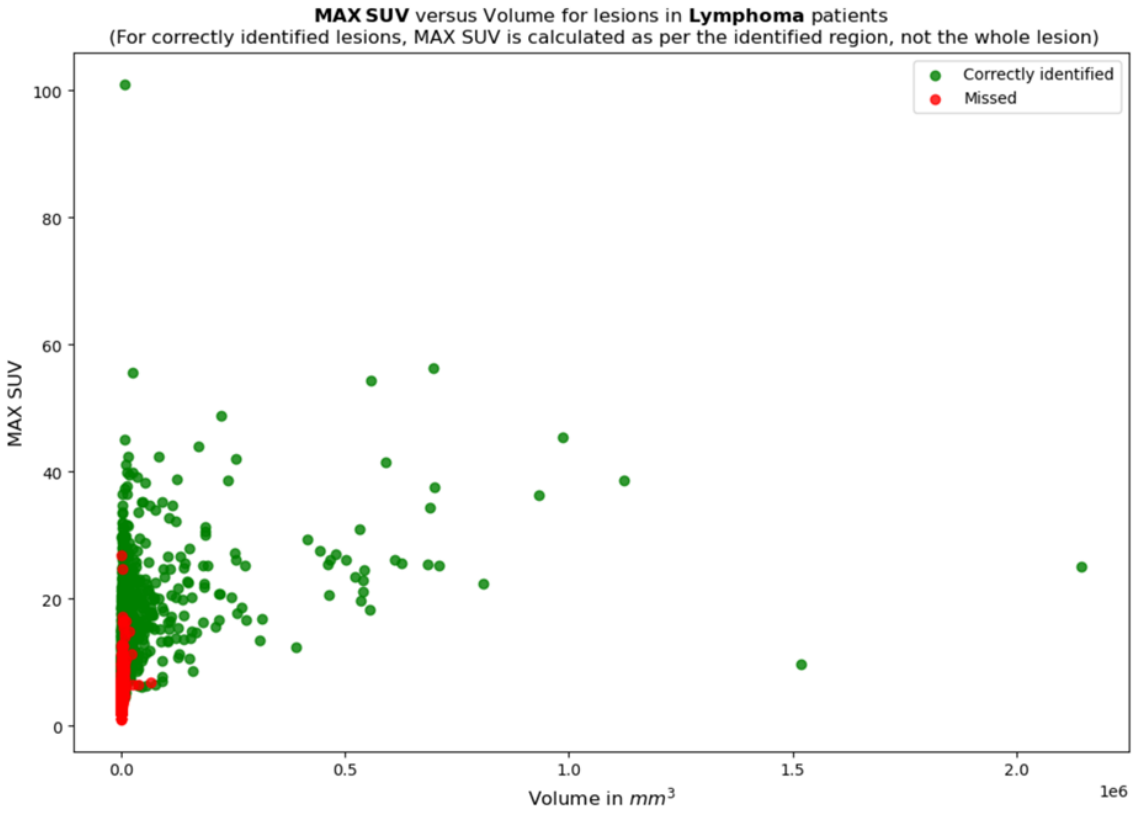

Figure A.18: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lymphoma patients.

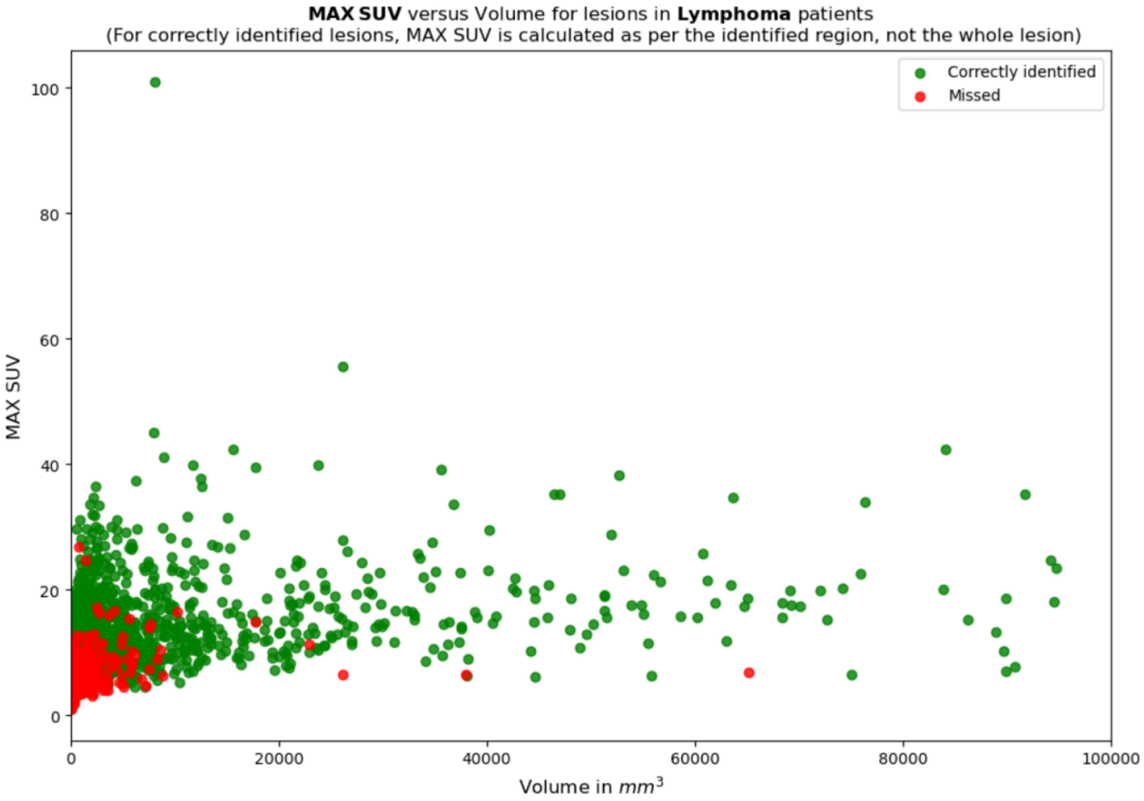

Figure A.19: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in lymphoma patients with limited x-axis for a better view on the low volume lesions.

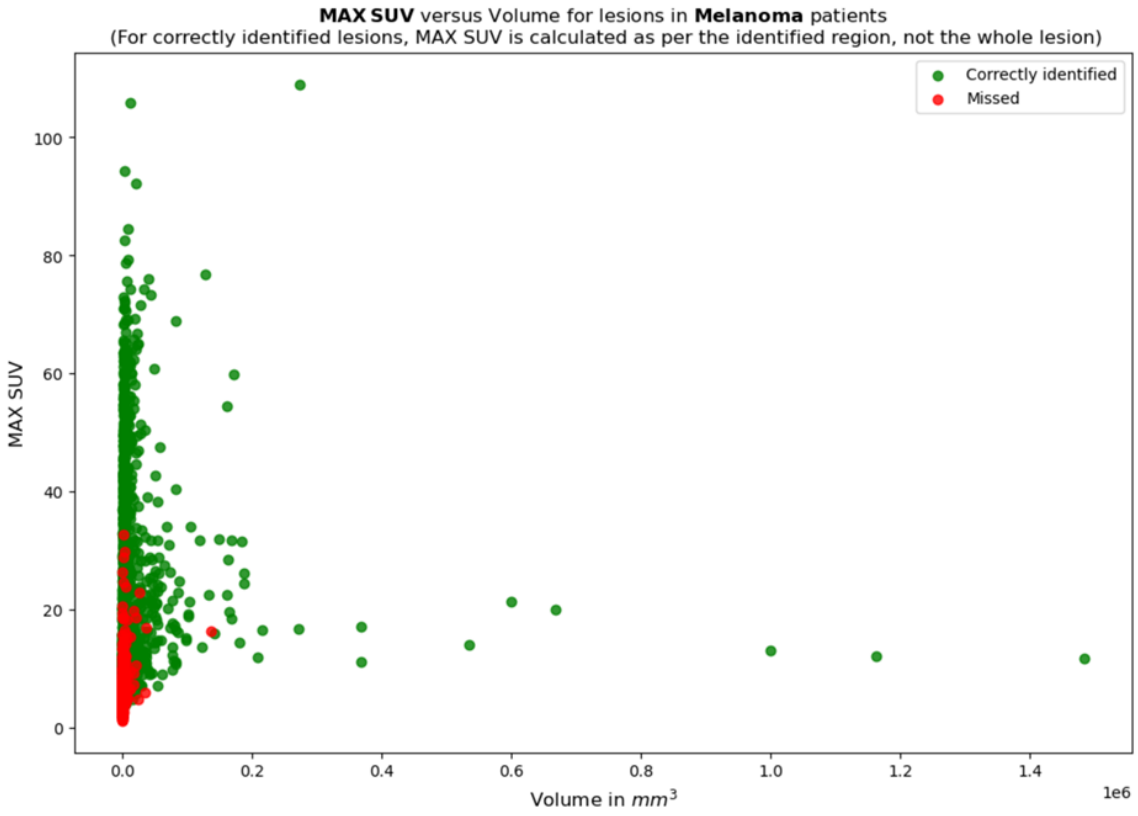

Figure A.20: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in melanoma patients.

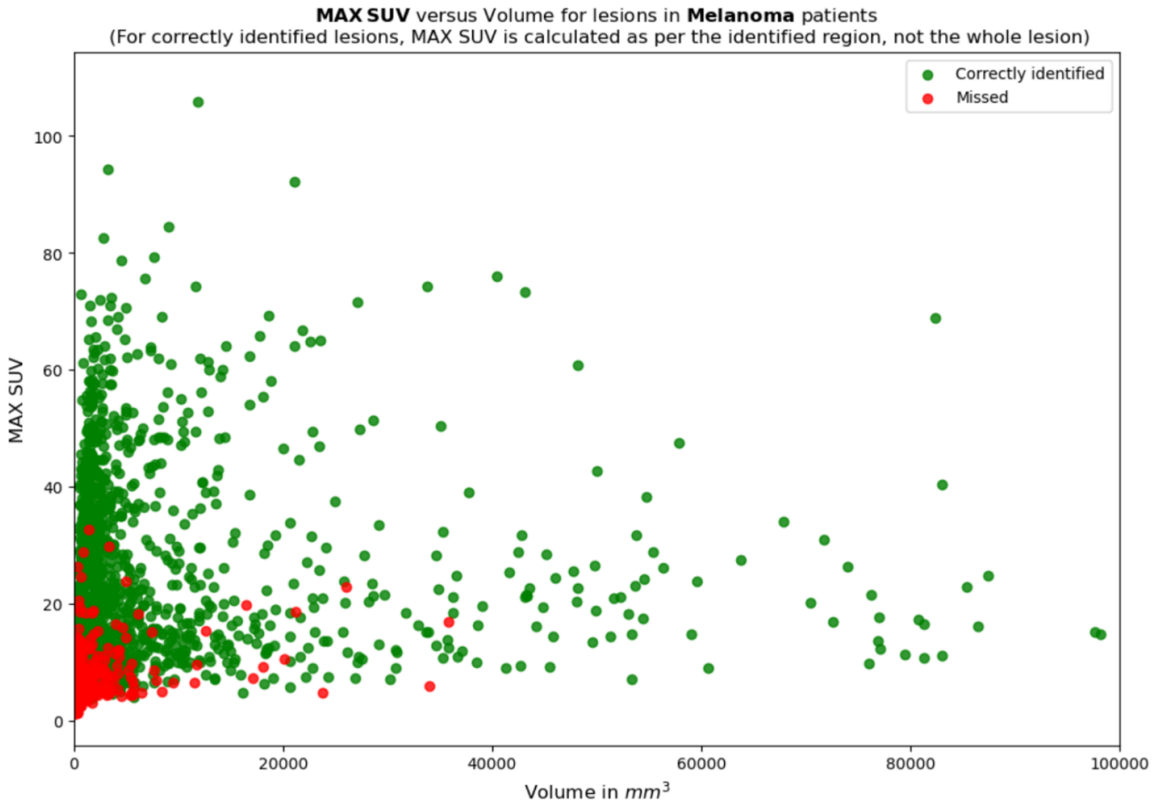

Figure A.21: Maximum SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, MAX SUV of the identified portion is reported. For the completely missed lesions, MAX SUV of the whole lesion is reported in melanoma patients with limited x-axis for a better view on the low volume lesions.

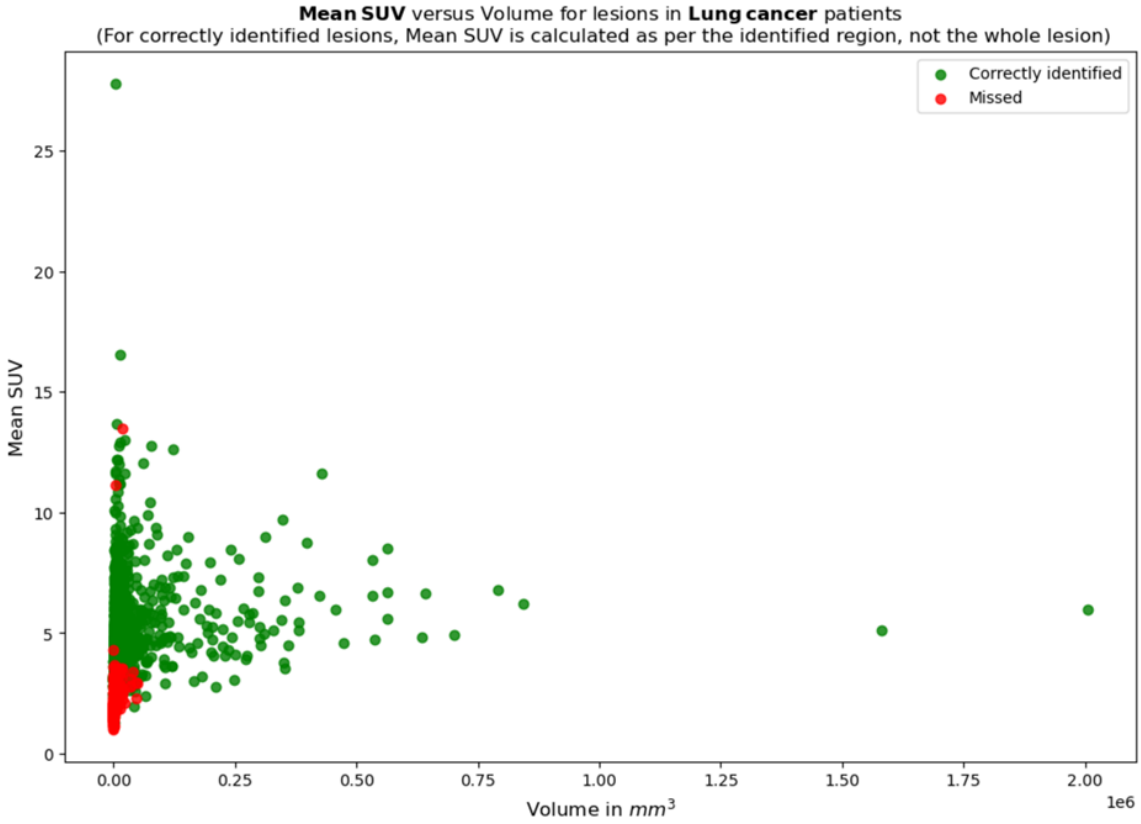

Figure A.22: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lung cancer patients.

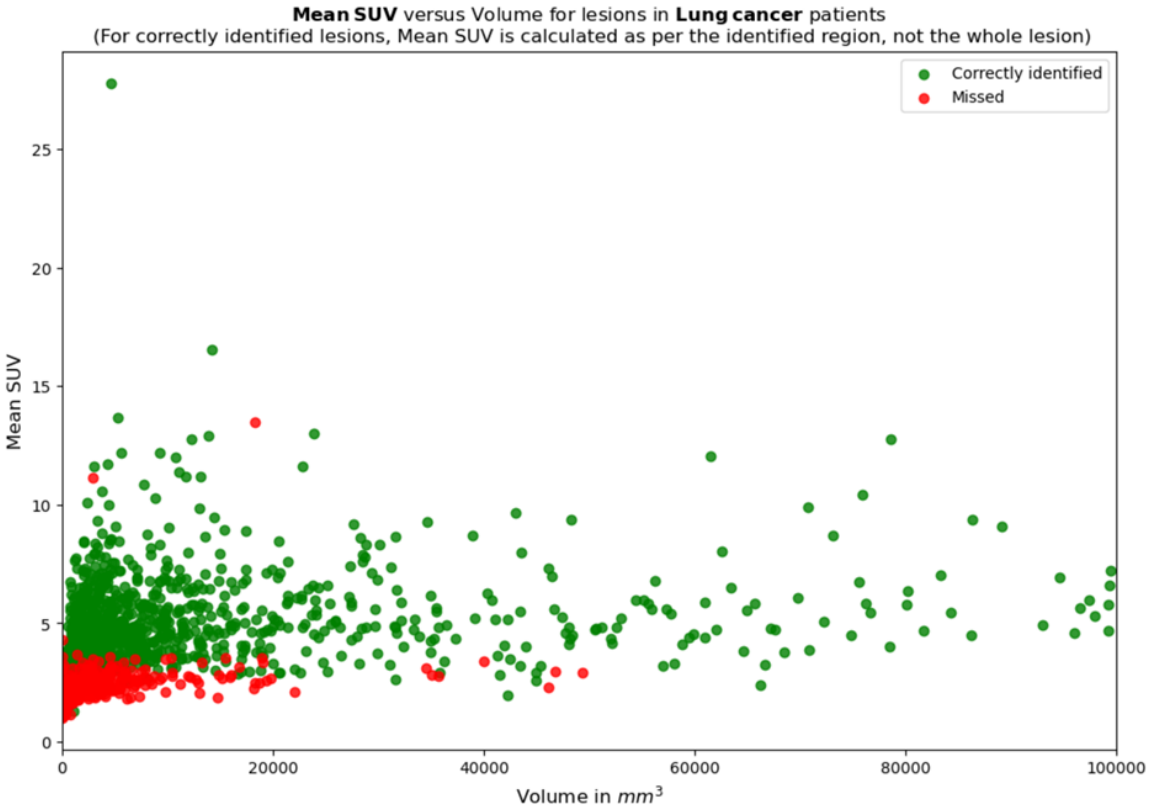

Figure A.23: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lung cancer patients with limited x-axis for a better view on the low volume lesions.

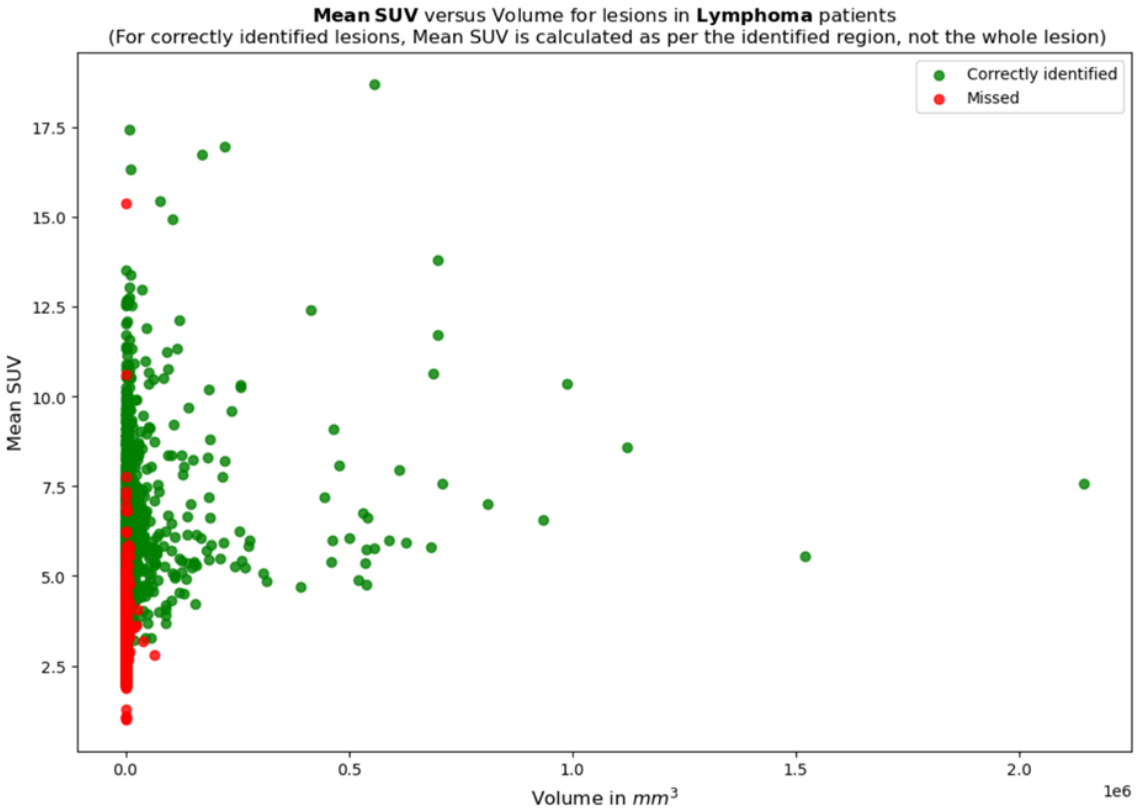

Figure A.24: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lymphoma patients.

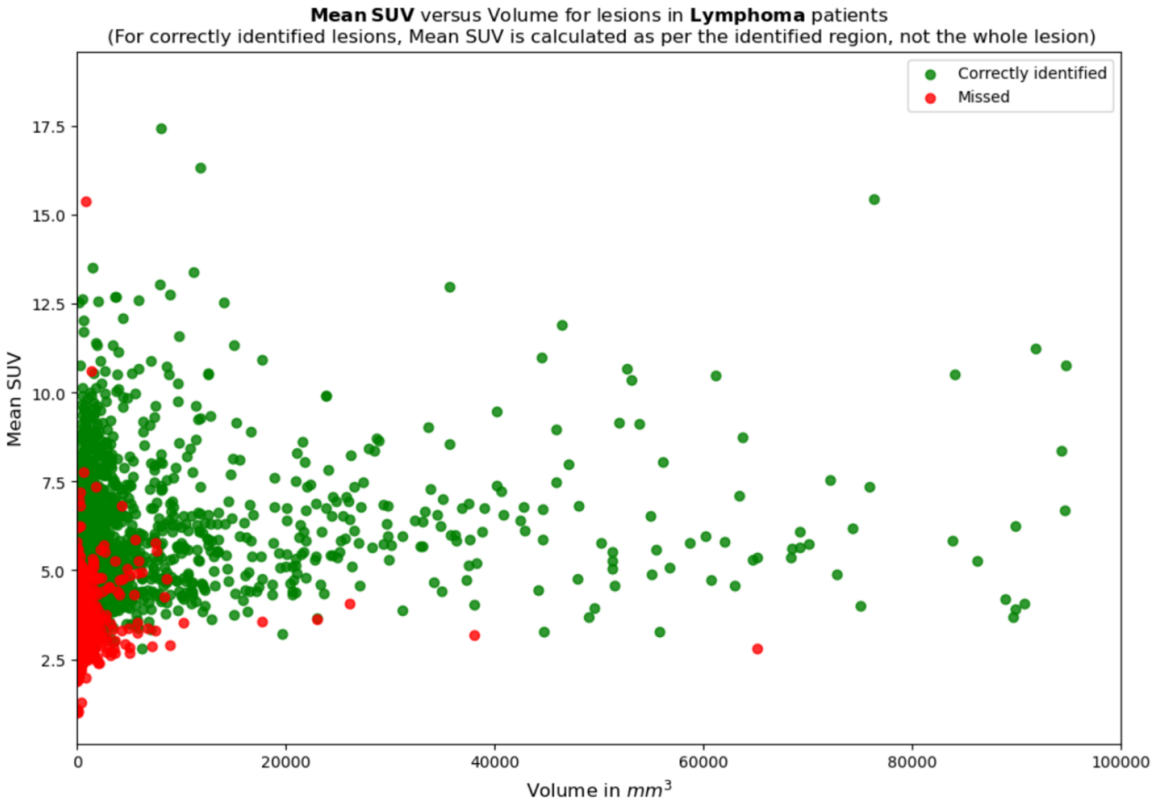

Figure A.25: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in lymphoma patients with limited x-axis for a better view on the low volume lesions.

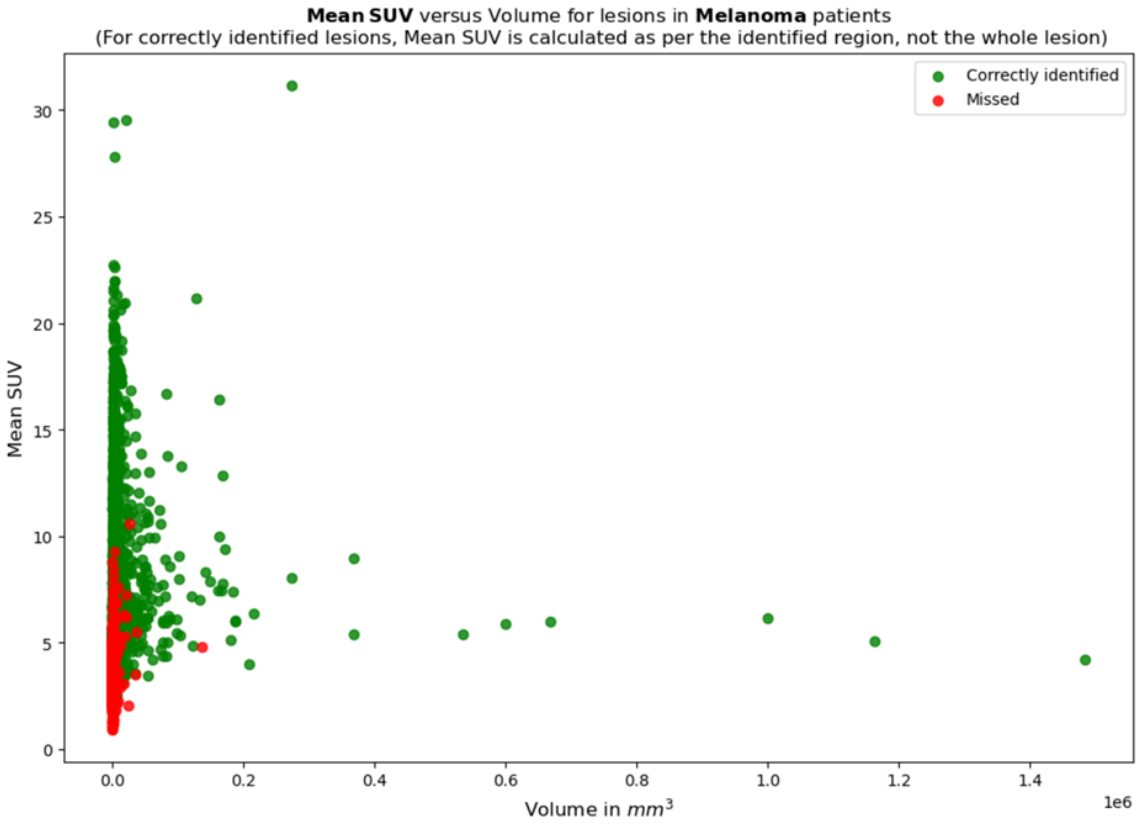

Figure A.26: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in melanoma patients.

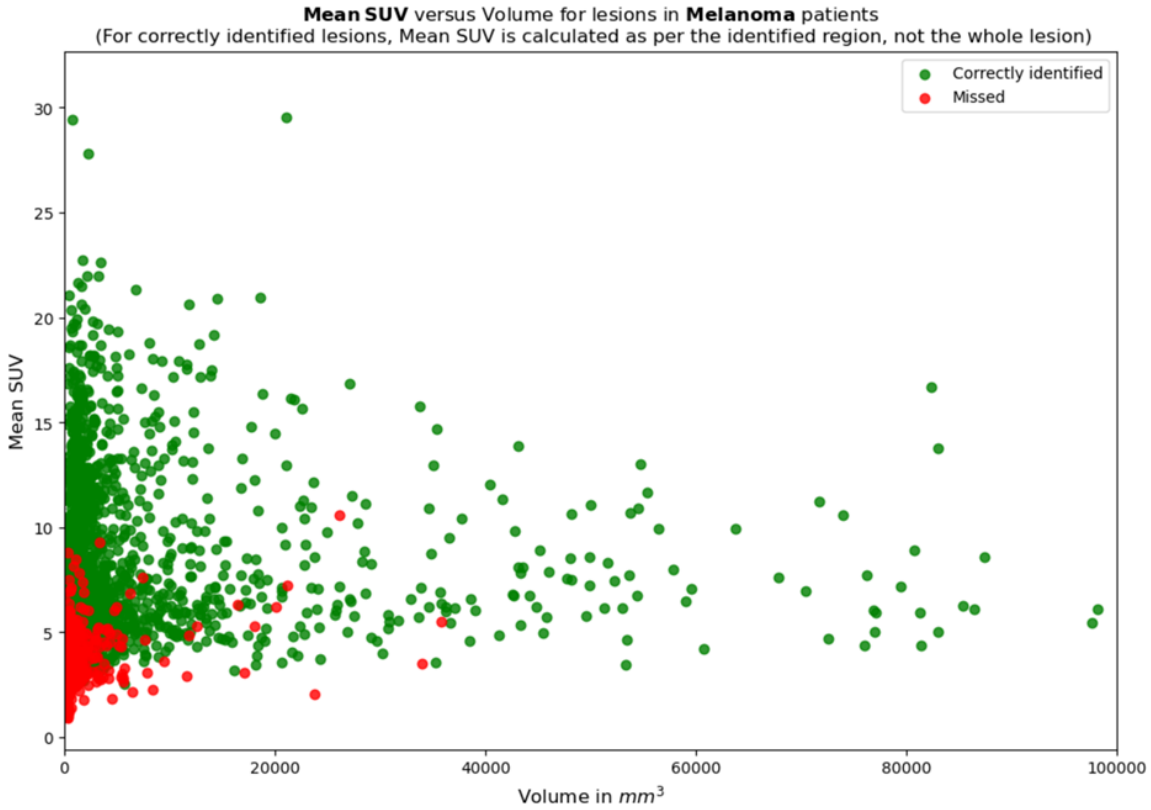

Figure A.27: Mean SUV captured in lesions. For lesions that were partially identified (segmented) by nnU-Net, Mean SUV of the identified portion is reported. For the completely missed lesions, Mean SUV of the whole lesion is reported in melanoma patients with limited x-axis for a better view on the low volume lesions.

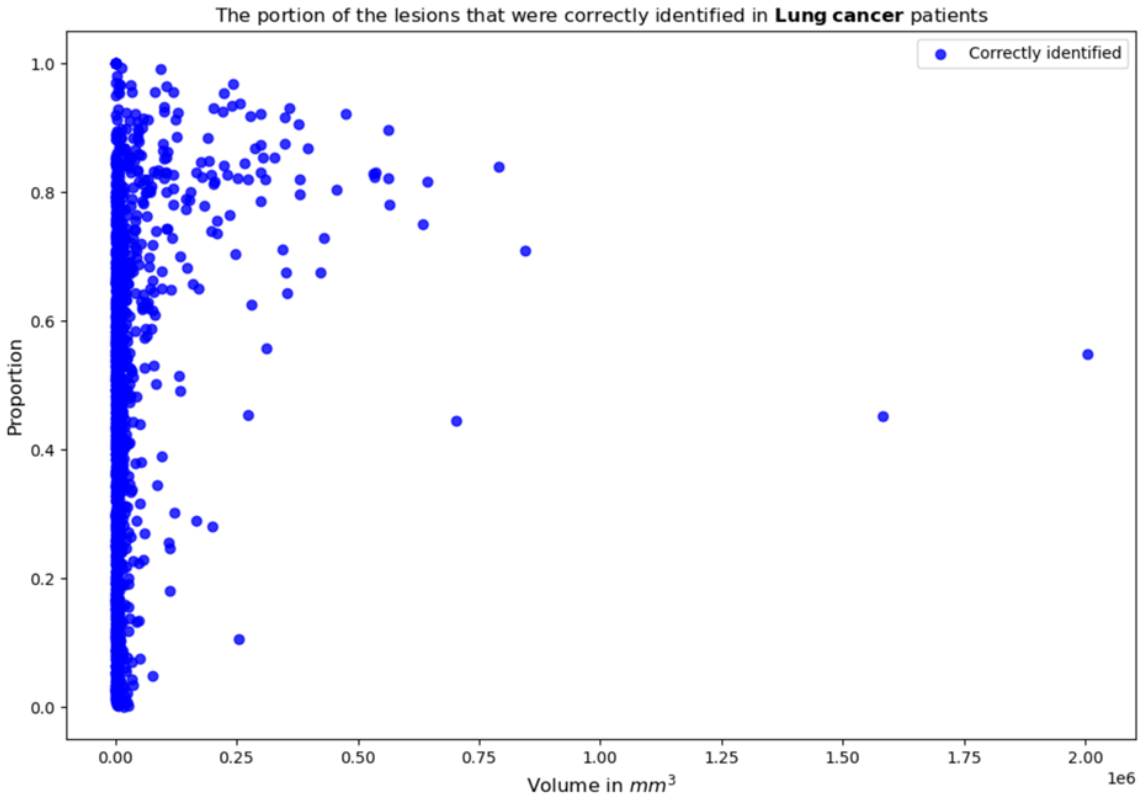

Figure A.28: Percentage of each lesion that has been correctly identified with respect to their volumes in lung cancer patients for nnU-Net algorithm.

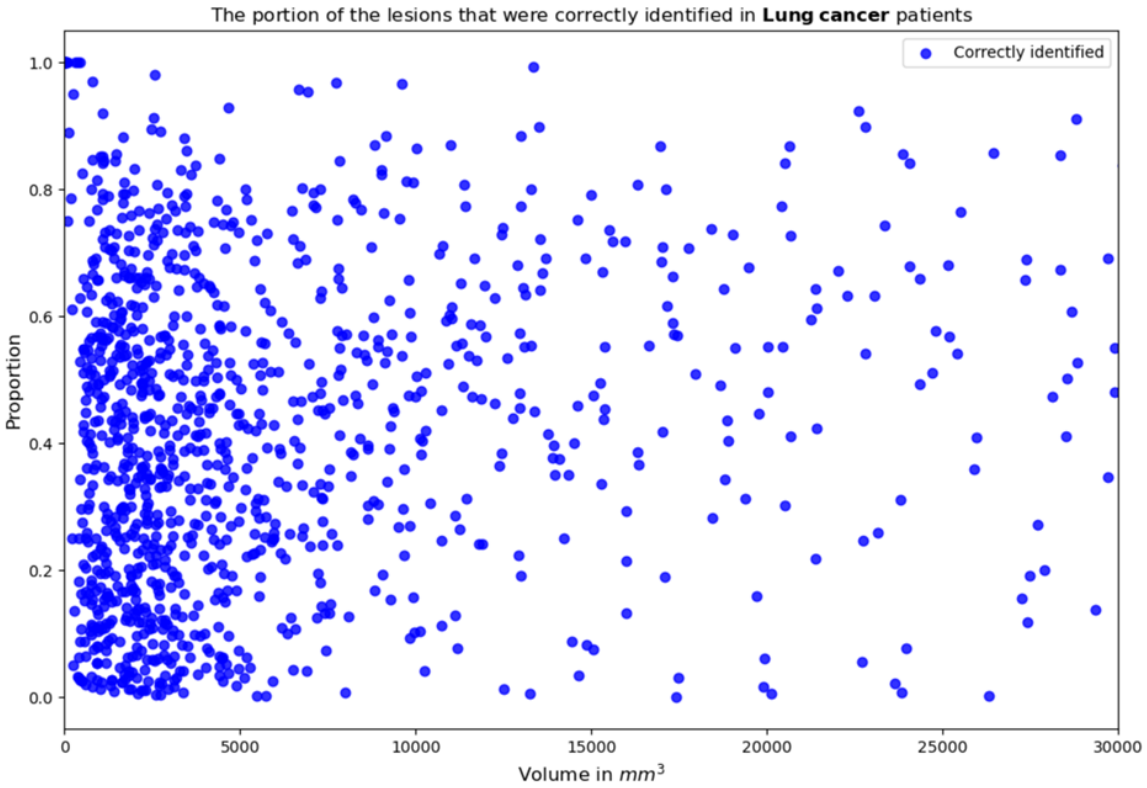

Figure A.29: Percentage of each lesion that has been correctly identified with respect to their volumes in lung cancer patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm.

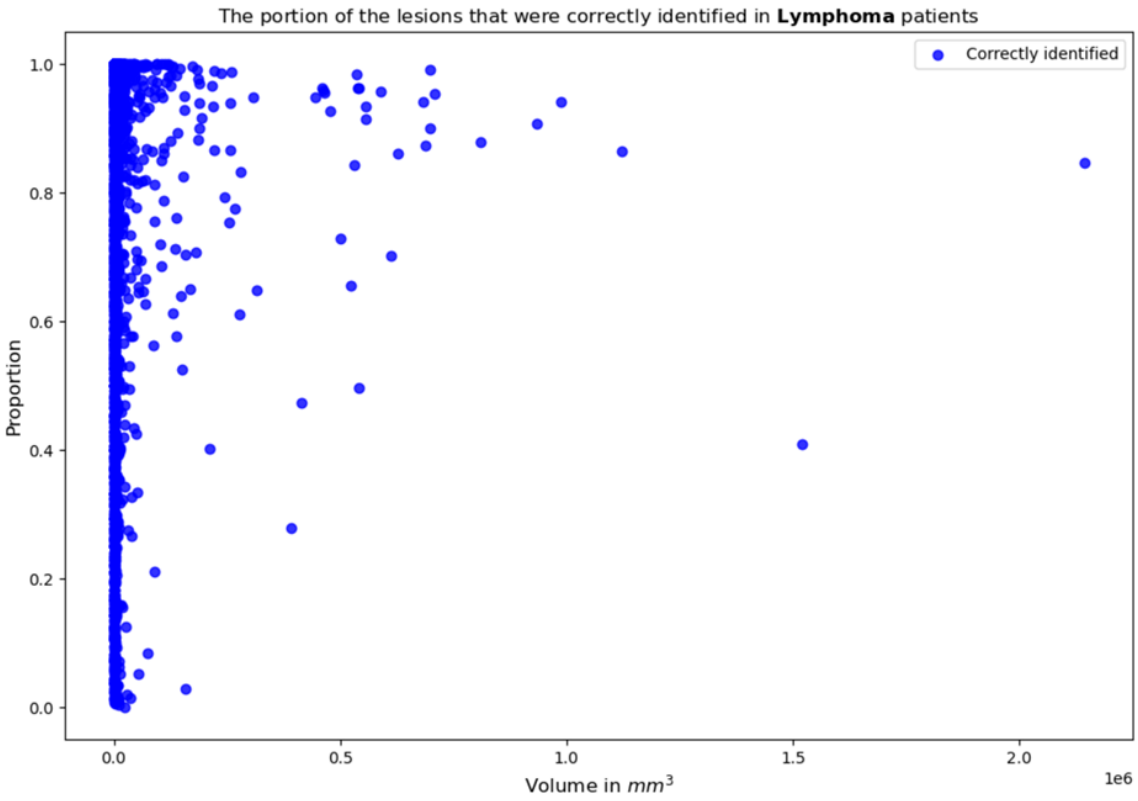

Figure A.30: Percentage of each lesion that has been correctly identified with respect to their volumes in lymphoma patients for nnU-Net algorithm.

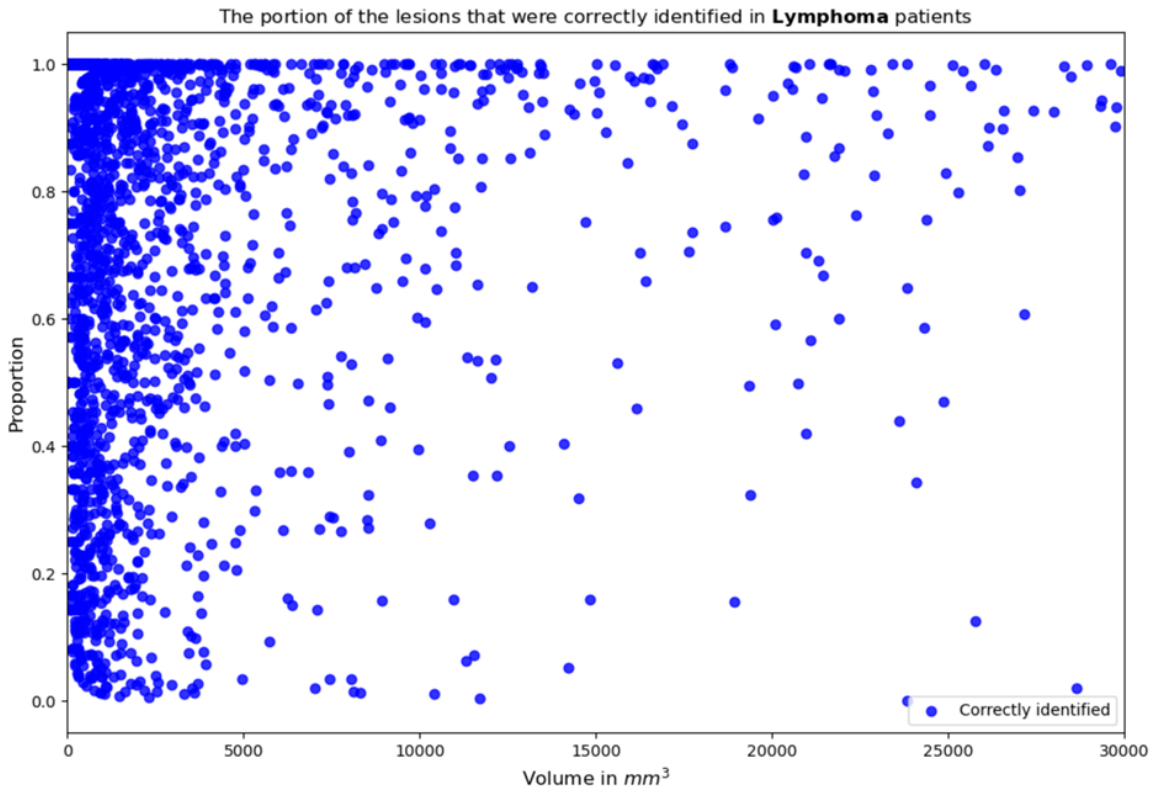

Figure A.31: Percentage of each lesion that has been correctly identified with respect to their volumes in lymphoma patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm.

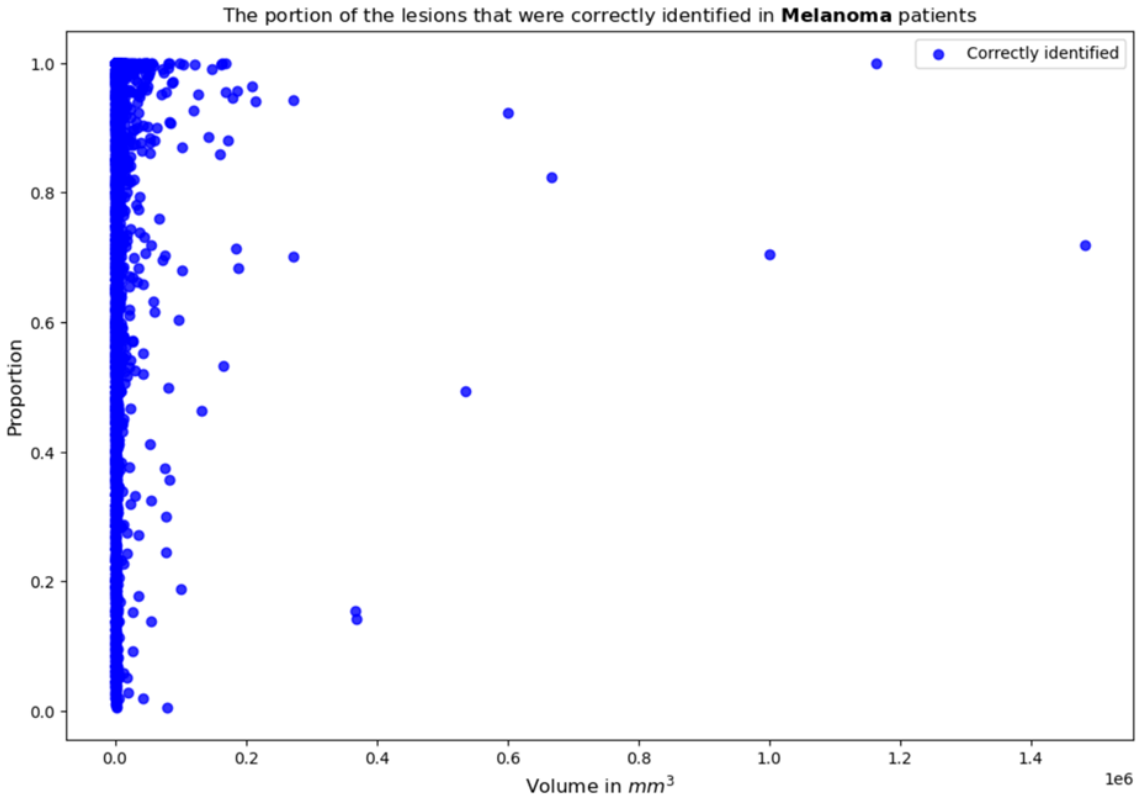

Figure A.32: Percentage of each lesion that has been correctly identified with respect to their volumes in melanoma patients for nnU-Net algorithm.

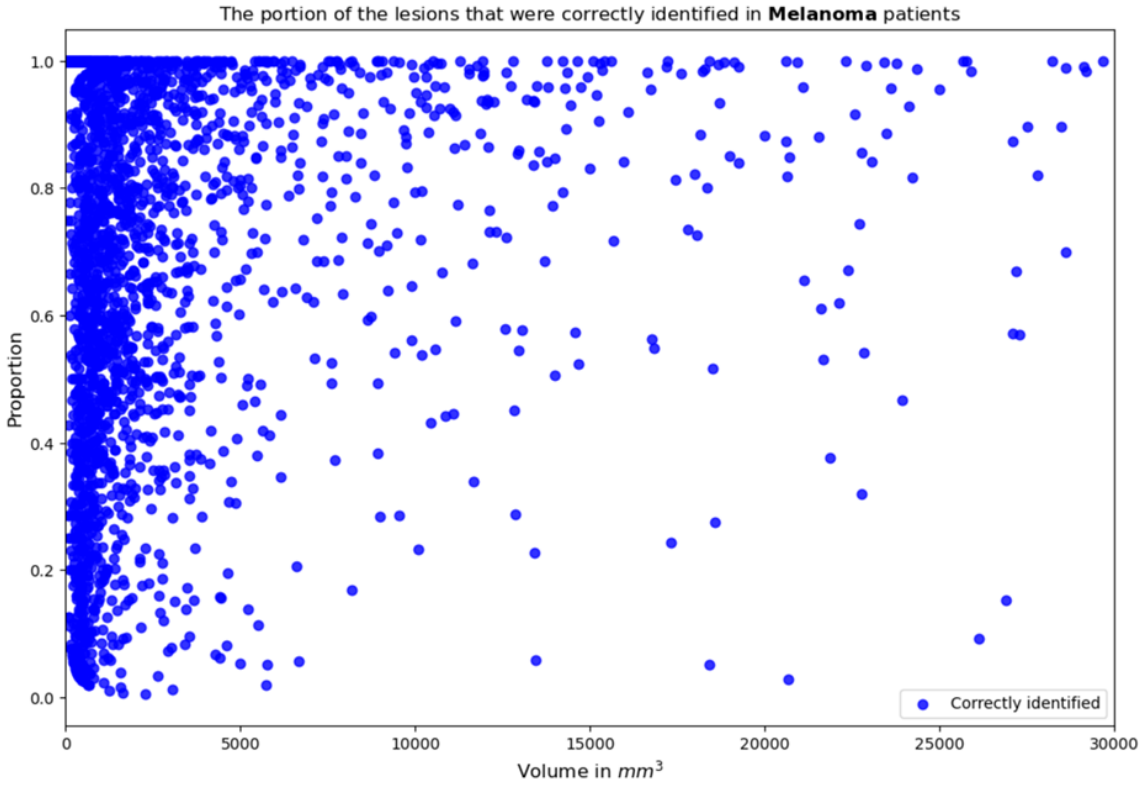

Figure A.33: Percentage of each lesion that has been correctly identified with respect to their volumes in melanoma patients with limited x-axis for a better view on the low volume lesions for nnU-Net algorithm.

Appendix B

B.1 Code for connected component analysis

```
1 import SimpleITK as sitk
2 import numpy as np
3 import os
4
5 def do_component_analysis(path, filename):
6     sitkimageobj = sitk.ReadImage(os.path.join(path, filename), sitk.
    sitkUInt8)
7     component_imageobj = sitk.ConnectedComponent(sitkimageobj)
8     sorted_component_imageobj = sitk.RelabelComponent(component_imageobj
    , sortByObjectSize=True)
9     imagearray = sitk.GetArrayFromImage(sorted_component_imageobj)
10    return np.transpose(imagearray, (1, 2, 0))
```

Listing B.1: Disconnected component analysis function

B.2 Code for random dilation

```
1 import random
```

```
import numpy as np
from scipy.ndimage import binary_dilation

def random_dilation(binary_image, min_size=1, max_size=5):
    # Generate random size for structuring element
    size_x = random.randint(min_size, max_size)
    size_y = random.randint(min_size, max_size)
    size_z = random.randint(min_size, max_size)

    # Create the random-sized structuring element
    structuring_element = np.ones((size_x, size_y, size_z), dtype=bool)

    # Perform dilation
    return binary_dilation(binary_image, structure=structuring_element)
```

Listing B.2: Random dilation function

B.3 Code for CT image pre-processing

```
import numpy as np

def preprocess_CT(image):
    WINDOW_LEVEL = 40
    WINDOW_WIDTH = 400
    lower_bound = WINDOW_LEVEL - WINDOW_WIDTH / 2
    upper_bound = WINDOW_LEVEL + WINDOW_WIDTH / 2
    image_preprocessed = np.clip(image, lower_bound, upper_bound)
    image_preprocessed = (
        (image_preprocessed - np.min(image_preprocessed))
```

```
11          / (np.max(image_preprocessed) - np.min(image_preprocessed))
12          * 255.0
13      )
14      return np.uint8(image_preprocessed)
```

Listing B.3: CT image pre-processing

B.4 Code for SUV contrast stretching

```
1 import numpy as np
2 import cv2
3
4 def contrast_stretching(img):
5     minmax_img = np.zeros_like(img)
6     for i in range(img.shape[2]):
7         minmax_img[:, :, i] = cv2.normalize(img[:, :, i], None, 0, 255,
      cv2.NORM_MINMAX)
8     return minmax_img.astype(np.uint8)
```

Listing B.4: SUV contrast stretching

Bibliography

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[2] D. E. Alvarado-Carrillo, I. Cruz-Aceves, M. A. Hernández-González, and L. M. López-Montero, "Robust detection and modeling of the major temporal arcade in retinal fundus images," *Mathematics*, vol. 10, no. 8, p. 1334, 2022.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] P. Gupta, "nnu-net : The no-new-unet for automatic segmentation,"

2020. [Online]. Available: https://medium.com/miccai-educational-initiative/nnu-net-the-no-new-unet-for-automatic-segmentation-8d655f3f6d2a

[6] I. Contributors, "3d binary filters," https://imagej.net/plugins/3d-binary-filters, 2024, accessed: 2024-06-11.

[7] Meta, "Segment anything github repository," https://github.com/facebookresearch/segment-anything, 2023.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[10] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[11] L. Fass, "Imaging and cancer: a review," *Molecular oncology*, vol. 2, no. 2, pp. 115–152, 2008.

[12] N. C. Institute, "Uses of imaging," https://imaging.cancer.gov/imaging_basics/cancer_imaging/uses_of_imaging.htm, 2024, accessed: 2024-06-04.

[13] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two

decades and the challenges ahead," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 85–106, 2000.

[14] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[15] A. Belle, R. Thiagarajan, S. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed research international*, vol. 2015, no. 1, p. 370194, 2015.

[16] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[17] D. J. Withey and Z. J. Koles, "Medical image segmentation: Methods and software," in *2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*. IEEE, 2007, pp. 140–143.

[18] M. Jena, S. P. Mishra, and D. Mishra, "A survey on applications of machine learning techniques for medical image segmentation," *Internationa Journal of Engineering & Technology*, vol. 7, no. 4, pp. 4489–4495, 2018.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image

segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[21] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical image analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[22] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[23] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[24] K. Li, Y. Fang, W. Li, C. Pan, P. Qin, Y. Zhong, X. Liu, M. Huang, Y. Liao, and S. Li, "Ct image visual quantitative evaluation and clinical classification of coronavirus disease (covid-19)," *European radiology*, vol. 30, pp. 4407–4416, 2020.

[25] R. Robb and C. Barillot, "Interactive display and analysis of 3-d medical images," *IEEE transactions on medical imaging*, vol. 8, no. 3, pp. 217–226, 1989.

[26] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[27] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics*, vol. 35, no. 1, pp. 3–14, 2010.

[28] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[29] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, pp. 1045–1057, 2013.

[30] J. W. Fletcher, B. Djulbegovic, H. P. Soares, B. A. Siegel, V. J. Lowe, G. H. Lyman, R. E. Coleman, R. Wahl, J. C. Paschold, N. Avril *et al.*, "Recommendations on the use of 18f-fdg pet in oncology," *Journal of Nuclear Medicine*, vol. 49, no. 3, pp. 480–508, 2008.

[31] S. Basu, T. C. Kwee, S. Surti, E. A. Akin, D. Yoo, and A. Alavi, "Fundamentals of pet and pet/ct imaging," *Annals of the New York Academy of Sciences*, vol. 1228, no. 1, pp. 1–18, 2011.

[32] P. E. Kinahan, D. Townsend, T. Beyer, and D. Sashin, "Attenuation correction for a combined 3d pet/ct scanner," *Medical physics*, vol. 25, no. 10, pp. 2046–2053, 1998.

[33] M. Motwani, "2022 artificial intelligence primer for the nuclear cardiologist," *Journal of Nuclear Cardiology*, vol. 30, no. 6, pp. 2441–2453, 2023.

[34] M. Fallahpoor, S. Chakraborty, B. Pradhan, O. Faust, P. D. Barua, H. Chegeni, and R. Acharya, "Deep learning techniques in pet/ct imaging: A comprehensive review from sinogram to image space," *Computer Methods and Programs in Biomedicine*, p. 107880, 2023.

[35] C. Lindner, C.-W. Wang, C.-T. Huang, C.-H. Li, S.-W. Chang, and T. F. Cootes, "Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms," *Scientific reports*, vol. 6, no. 1, p. 33581, 2016.

[36] M. D. Farwell, D. A. Pryma, and D. A. Mankoff, "Pet/ct imaging in cancer: current applications and future directions," *Cancer*, vol. 120, no. 22, pp. 3433–3445, 2014.

[37] D. F. Santos, M. E. Takahashi, M. Camacho, M. d. C. L. de Lima, B. J. Amorim, E. M. Rohren, and E. Etchebehere, "Whole-body tumor burden in pet/ct expert review," *Clinical and Translational Imaging*, vol. 11, no. 1, pp. 5–22, 2023.

[38] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[40] R. E Woods and R. C Gonzalez, "Digital image processing," 2008.

[41] S. Capitanio, A. J. Nordin, A. R. Noraini, and C. Rossetti, "Pet/ct in nononcological lung diseases: current applications and future perspectives," *European Respiratory Review*, vol. 25, no. 141, pp. 247–258, 2016.

[42] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, "A review on segmentation of positron emission tomography images," *Computers in biology and medicine*, vol. 50, pp. 76–96, 2014.

[43] D. Han, J. Bayouth, Q. Song, A. Taurani, M. Sonka, J. Buatti, and X. Wu, "Globally optimal tumor segmentation in pet-ct images: a graph-based co-segmentation method," in *Information Processing in Medical Imaging: 22nd International Conference, IPMI 2011, Kloster Irsee, Germany, July 3-8, 2011. Proceedings 22*. Springer, 2011, pp. 245–256.

[44] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 243–254, 2016.

[45] H. Tian, T. Zhu, W. Liu, and W. Zhou, "Image fairness in deep learning: problems, models, and challenges," *Neural Computing and Applications*, vol. 34, no. 15, pp. 12 875–12 893, 2022.

[46] K. Ikushima, H. Arimura, Z. Jin, H. Yabu-Uchi, J. Kuwazuru, Y. Shioyama, T. Sasaki, H. Honda, and M. Sasaki, "Computer-assisted framework for machine-learning–based delineation of gtv regions on datasets of planning ct and pet/ct images," *Journal of radiation research*, vol. 58, no. 1, pp. 123–134, 2017.

[47] Z. Jin, H. Arimura, Y. Shioyama, K. Nakamura, J. Kuwazuru, T. Magome, H. Yabu-Uchi, H. Honda, H. Hirata, and M. Sasaki, "Computer-assisted delineation of lung tumor regions in treatment planning ct images with pet/ct image

sets based on an optimum contour selection method," *Journal of radiation research*, vol. 55, no. 6, pp. 1153–1162, 2014.

[48] E. Grossiord, H. Talbot, N. Passat, M. Meignan, and L. Najman, "Automated 3d lymphoma lesion segmentation from pet/ct characteristics," in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE, 2017, pp. 174–178.

[49] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[50] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[52] 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.

[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[54] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[55] L. Zhao and Z. Zhang, "A improved pooling method for convolutional neural networks," *Scientific Reports*, vol. 14, no. 1, p. 1589, 2024.

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[57] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 803–23 828.

[58] L. Xu, G. Tetteh, J. Lipkova, Y. Zhao, H. Li, P. Christ, M. Piraud, A. Buck, K. Shi, B. H. Menze *et al.*, "Automated whole-body bone lesion detection for multiple myeloma on 68 ga-pentixafor pet/ct imaging using deep learning methods," *Contrast media & molecular imaging*, vol. 2018, 2018.

[59] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[60] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[61] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3507–3516, 2021.

[62] Z. Zhong, Y. Kim, L. Zhou, K. Plichta, B. Allen, J. Buatti, and X. Wu, "3d fully convolutional networks for co-segmentation of tumors on pet-ct images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 228–231.

[63] I. Dirks, M. Keyaerts, B. Neyns, and J. Vandemeulebroucke, "Computer-aided detection and segmentation of malignant melanoma lesions on whole-body 18f-fdg pet/ct using an interpretable deep learning approach," *Computer methods and programs in biomedicine*, vol. 221, p. 106902, 2022.

[64] B. Huang, Z. Chen, P.-M. Wu, Y. Ye, S.-T. Feng, C.-Y. O. Wong, L. Zheng, Y. Liu, T. Wang, Q. Li *et al.*, "Fully automated delineation of gross tumor volume for head and neck cancer on pet-ct using deep learning: a dual-center study," *Contrast media & molecular imaging*, vol. 2018, 2018.

[65] X. Zhao, L. Li, W. Lu, and S. Tan, "Tumor co-segmentation in pet/ct using multi-modality fully convolutional neural network," *Physics in Medicine & Biology*, vol. 64, no. 1, p. 015011, 2018.

[66] D. Xiang, B. Zhang, Y. Lu, and S. Deng, "Modality-specific segmentation network for lung tumor segmentation in pet-ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1237–1248, 2022.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[68] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps

from pet-ct images of lung cancer," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 204–217, 2019.

[69] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, M. Vallieres, S. Zhu, J. Xie, Y. Peng *et al.*, "Head and neck tumor segmentation in pet/ct: the hecktor challenge," *Medical image analysis*, vol. 77, p. 102336, 2022.

[70] A. Iantsen, D. Visvikis, and M. Hatt, "Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined pet and ct images," in *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*. Springer, 2021, pp. 37–43.

[71] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[72] J. Ma and X. Yang, "Combining cnn and hybrid active contours for head and neck tumor segmentation in ct and pet images," in *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*. Springer, 2021, pp. 59–64.

[73] S. Jemaa, J. Fredrickson, R. A. Carano, T. Nielsen, A. de Crespigny, and T. Bengtsson, "Tumor segmentation and feature extraction from whole-body

fdg-pet/ct using cascaded 2d and 3d convolutional neural networks," *Journal of digital imaging*, vol. 33, pp. 888–894, 2020.

[74] A. R. Groendahl, I. S. Knudtsen, B. N. Huynh, M. Mulstad, Y. M. Moe, F. Knuth, O. Tomic, U. G. Indahl, T. Torheim, E. Dale *et al.*, "A comparison of methods for fully automatic segmentation of tumors and involved nodes in pet/ct of head and neck cancers," *Physics in Medicine & Biology*, vol. 66, no. 6, p. 065012, 2021.

[75] Z. Zhong, Y. Kim, K. Plichta, B. G. Allen, L. Zhou, J. Buatti, and X. Wu, "Simultaneous cosegmentation of tumors in pet-ct images using deep fully convolutional networks," *Medical physics*, vol. 46, no. 2, pp. 619–633, 2019.

[76] J. Xie and Y. Peng, "The head and neck tumor segmentation using nnu-net with spatial and channel 'squeeze & excitation'blocks," in *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*. Springer, 2021, pp. 28–36.

[77] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[78] L. Li, X. Zhao, W. Lu, and S. Tan, "Deep learning for variational multimodality tumor segmentation in pet/ct," *Neurocomputing*, vol. 392, pp. 277–295, 2020.

[79] L. Sibille, X. Zhan, and L. Xiang, "Whole-body tumor segmentation of 18f-fdg

pet/ct using a cascaded and ensembled convolutional neural networks," *arXiv preprint arXiv:2210.08068*, 2022.

[80] Y. Peng, J. Kim, D. Feng, and L. Bi, "Automatic tumor segmentation via false positive reduction network for whole-body multi-modal pet/ct images," *arXiv preprint arXiv:2209.07705*, 2022.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[82] A. G. Challenge, "Autopets grand challenge," https://autopet.grand-challenge.org/, 2022, accessed: 2024-06-29.

[83] N. E. Protonotarios, I. Katsamenis, S. Sykiotis, N. Dikaios, G. A. Kastis, S. N. Chatziioannou, M. Metaxas, N. Doulamis, and A. Doulamis, "A few-shot u-net deep learning model for lung cancer lesion segmentation via pet/ct imaging," *Biomedical Physics & Engineering Express*, vol. 8, no. 2, p. 025019, 2022.

[84] Z. Huang, S. Tang, Z. Chen, G. Wang, H. Shen, Y. Zhou, H. Wang, W. Fan, D. Liang, Y. Hu *et al.*, "Tg-net: Combining transformer and gan for nasopharyngeal carcinoma tumor segmentation based on total-body uexplorer pet/ct scanner," *Computers in Biology and Medicine*, vol. 148, p. 105869, 2022.

[85] L. Huang, S. Ruan, P. Decazes, and T. Denœux, "Lymphoma segmentation from 3d pet-ct images using a deep evidential network," *International Journal of Approximate Reasoning*, vol. 149, pp. 39–60, 2022.

[86] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[87] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 656–666.

[88] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[89] S. Bionetworks, "Synapse multi-organ ct dataset," 2015, accessed: 2024-07-01. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[90] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran, and D. Rubin, "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions," *Scientific Data*, vol. 9, no. 1, p. 601, 2022.

[91] lab-midas, "Tcia_processing," https://github.com/lab-midas/TCIA_processing, 2024, autoPET TCIA pipeline.

[92] I. Segmentation and R. T. (ITK), "Assign contiguous labels to connected regions in an image," https://examples.itk.org/src/segmentation/

connectedcomponents/assigncontiguouslabelstoconnectedregions/ documentation, 2023, iTK Examples.

[93] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[94] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[95] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[96] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International workshop on machine learning in medical imaging*. Springer, 2017, pp. 379–387.

[97] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

[98] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[99] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[100] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[101] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[102] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[103] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li, "Sam on medical images: A comprehensive study on three prompt modes," *arXiv preprint arXiv:2305.00035*, 2023.

[104] J. Bai, F. Lu, K. Zhang *et al.*, "Onnx: Open neural network exchange," https://github.com/onnx/onnx, 2019.

[105] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[106] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[107] A. Alloula, D. R. McGowan, and B. W. Papież, "Autopet challenge 2023: nnunet-based whole-body 3d pet-ct tumour segmentation," *arXiv preprint arXiv:2309.13675*, 2023.

[108] S. Gatidis, M. Früh, M. Fabritius, S. Gu, K. Nikolaou, C. La Fougère, J. Ye, J. He, Y. Peng, L. Bi *et al.*, "The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging," *Research Square*, 2023.

[109] Lightning AI, "Gradient accumulation: How to train with large batch sizes," 2024, accessed: 2024-06-27. [Online]. Available: https://lightning.ai/blog/gradient-accumulation/

[110] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[111] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for

semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[112] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomedical Signal Processing and Control*, vol. 84, p. 104791, 2023.