NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

DEPARTMENT OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

Development and Evaluation of an Interactive Machine Learning-based Method for Segmentation of Solid Tumors and Organs

DIPLOMA THESIS

Dimitrios A. Bounias

Supervisor : Konstantina S. Nikita

Professor at N.T.U.A.

Athens, October 2020

Page left intentionally blank.

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

DEPARTMENT OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

Development and Evaluation of an Interactive Machine Learning-based Method for Segmentation of Solid Tumors and Organs

DIPLOMA THESIS

Dimitrios A. Bounias

Supervisor : Konstantina S. Nikita

Professor at N.T.U.A.

Approved by the three-member examination committee on October 13, 2020.

.....	.....	.....
K. Nikita	G. Stamou	A. Stafylopatis
Professor at N.T.U.A.	Assoc. Professor at N.T.U.A.	Professor at N.T.U.A.

Athens, October 2020

.....

Dimitrios A. Bounias

Electrical and Computer Engineering Graduate, NTUA

Abstract

In this thesis, a fast, accurate, and consistent method for general-purpose medical image segmentation is presented. The method is based on interactive machine learning (IML).

Utilizing brief user training annotations and the adaptive geodesic distance transform, an ensemble of Support Vector Machines (SVMs) is trained, providing a patient-specific model applied to the whole image. To validate the method, retrospective cohorts were identified comprising 20 brain, 50 breast, and 50 lung cancer patients, as well as 20 spleen scans. Corresponding ground truth annotations were available for each scan. Two experts were recruited to segment each cohort twice with the IML method and twice manually. Evaluation endpoints included speed, spatial overlap, and consistency. Statistical significance was evaluated through the Wilcoxon signed-rank non-parametric test.

The IML method was faster than manual annotation in all structures, by 53.1% on average. A significant ($p<0.001$) overlap difference was found for spleen ($Dice_{IML}/Dice_{Manual}$=0.91/0.87), breast tumors ($Dice_{IML}/Dice_{Manual}$=0.84/0.82), and lung nodules ($Dice_{IML}/Dice_{Manual}$=0.78/0.83). For intra-rater consistency, a significant ($p=0.003$) and favorable difference was found only for spleen ($Dice_{IML}/Dice_{Manual}$=0.91/0.89). For inter-rater consistency significant ($p<0.045$) differences were found for spleen ($Dice_{IML}/Dice_{Manual}$=0.91/0.87), breast ($Dice_{IML}/Dice_{Manual}$=0.86/0.81), lung ($Dice_{IML}/Dice_{Manual}$=0.85/0.89), the non-enhancing brain tumor sub-region ($Dice_{IML}/Dice_{Manual}$=0.79/0.67), and the enhancing brain tumor sub-region ($Dice_{IML}/Dice_{Manual}$=0.79/0.84), which in aggregation favored the IML algorithm.

The quantitative evaluation for speed, spatial overlap, and consistency, revealed the benefits of the proposed IML method when compared with manual annotation, for several clinically relevant problems. The IML method is publicly released through the Cancer Imaging Phenomics Toolkit (CaPTk - https://www.cbica.upenn.edu/captk) and as an MITK plugin (https://github.com/CBICA/InteractiveSegmentation).

Keywords

Machine learning, Support Vector Machines, C++, Image segmentation, Medical imaging, Magnetic Resonance Imaging, Computational tomography, Glioblastoma, breast cancer, lung cancer, Spleen segmentation

Acknowledgments

I would like to thank everyone who contributed to this diploma thesis, especially Professor Konstantina Nikita and Professor Christos Davatzikos, who gave me the opportunity to work on a topic that interests me a lot. They enabled me to delve deeper into the subject and gain valuable experience on how academia works. I also really appreciate that I was given the opportunity to live abroad for six months and work with people from various countries because of this collaboration.

I would like to also thank everyone with whom I worked with in the CBICA lab of the University of Pennsylvania during my stay in the US. The workplace climate was great and people working there were excellent, both as coworkers and as individuals. By working in CBICA, I was able to gain the necessary knowledge, experience, and support to prepare this diploma thesis, as well as to present it at the RSNA 2019 conference in Chicago.

Lastly, I also want to thank my family and friends who supported me and helped me throughout my studies.

Table of contents

List of figures

List of tables

1. Introduction

Medical image segmentation, i.e., annotation of regions of interest such as the area where a tumor resides, is an important task in clinical and research environments [1]–[5], facilitating subsequent computational analyses, which depend on the accuracy of the segmentation given as input [6]–[8]. Manual expert annotations are currently considered the gold standard, but tend to be tedious, time-consuming, and the results often have limited reproducibility [3], [9]. Even with the assistance of various tools [4], manual annotation can take multiple hours for a large scan or complex anatomies.

A plethora of fully automatic machine learning (ML) methods that can achieve state-of-the-art results have been proposed, but tend to face various challenges [10] that hinder clinical translation. Some of the most important challenges are generalization to unseen datasets and need for extensive expert corrections and refinements [4], [11].

Interactive Machine Learning (IML) methods fill the void between manual and automatic approaches by allowing an operator to train a patient-specific model via quick and rough drawings, which then automatically segments the entire scan [5], [12]–[14]. IML approaches provide the option for expedited refinements, and the final segmentation tends to get closer to the desired result as a function of the invested time. Two popular tools offering such functionality are ITK-SNAP [11] and 3D Slicer [15].

In this thesis, an IML method is proposed, leveraging adaptive geodesic distance (AGD) [16] maps alongside an ensemble of support vector machines (SVMs) that is agnostic to image type/dimensionality and supports multiparametric input images. The performance of the proposed method against manual expert segmentation was systematically evaluated across different anatomical structures and image modalities. Evaluation endpoints comprised speed, spatial overlap agreement, and consistency between different time-points and raters.

2. Introductory concepts

2.1. Radiology

Radiology is a multidisciplinary medical discipline devoted to the analysis of medical images for use in the diagnosis and treatment of diseases [17]. Examples of technologies used to acquire the medical images are MRI (Magnetic Resonance Imaging), CT (Computed Tomography), PET (Positron Emission Tomography), X-Ray, and ultrasound. As technologies advance and new ones enter the field rapidly, subspecialization is increasing in prevalence [18], although general radiologists are still common.

Diagnostic radiology is a field that focuses on using medical images for the purpose of diagnosing the causes of patient symptoms, screening for different diseases, and monitoring treatment response [19]. Diagnostic radiology is an important field in medicine, as it provides a non-interventional view of the body, which can be invaluable in assessing the health of a patient, with virtually no associated risk.

Interventional radiology focuses on assisting and guiding minor procedures undertaken in the human body, substituting the need for surgery. The practice can be used to insert instruments into the body, such as a catheter [19], and for cleaning blockages in the arteries [20]. Other examples of interventional radiology include needle biopsies of organs, embolization to control bleeding, and angiography.

2.2. Medical imaging concepts

Medical imaging is the process of noninvasively producing images of the interior of the body for clinical purposes. Medical images are typically greyscale and have some distinct properties that differentiate them from traditional images.

An important distinction of medical images is that there are a lot of different technologies, known as *modalities*, that are used to acquire them. Each modality produces results that look substantially different than the others and serves different purposes. Two well known modalities, that are used in the experiments hereafter, are MRI (Magnetic Resonance Imaging) and CT (Computed Tomography). MR scanners use radio waves and can achieve a high level of detail. CT scans are produced with x-rays, are typically less expensive, and are preferred for imaging organs and bones.

A multitude of acquisition protocols exist for each modality, each having different clinical interests. Images captured with the same modality but different acquisition protocols can be acquired during the same scan of the patient, using the same machine. For example, in the experiments regarding patients with glioblastoma, four MRI acquisition protocols were used: T1 and T2, which are standard protocols that mostly highlight fat and fluids respectively, T1CE (T1Gd) which contains information regarding angiogenesis and the integrity of the blood–brain barrier in the tumor and requires administering a contrast agent to the patient, and FLAIR which highlights the tumor and the peritumoral edematous region [6], [21]. Images taken with different acquisition protocols can be aligned to follow the same coordinate system through a process called "*registration*". As a result, these *co-registered* images can be processed as one image that has multiple channels, as it typically happens with the R/G/B channels of normal color images, enabling the application of various common computer vision methodologies.

A medical image usually has 2 or 3 dimensions. The 3 dimensions of a 3D image, define 3 planes when paired in groups of 2. These 3 planes are named sagittal, coronal, and axial. Visual inspection of 3D images can happen in a 2-dimensional monitor, by displaying 2D slices belonging to one of the three aforementioned planes. Medical images can also be 4-dimensional, where the fourth dimension is time, and are displayed in the same way as 3D

images, with an additional slider that controls time progression. 4D images are essentially a series of 3D images acquired over a small period of time and are clinically useful for inspecting how well a biological mechanism, like blood flow or oxygenation, operates in certain areas of the patient.

Segmentations are a special case of medical images used to delineate regions of interest in a scan. They are typically *co-registered* with the images they are describing and their pixels/voxels have 0 as a value everywhere except for the pixels/voxels covering the regions of interest, which have a distinct integer value for each region.

One other important aspect of medical images is that the physical distance between voxels, commonly referred to as "*spacing*", can be variable in each dimension. That is required because scans are acquired in 2D *slices*, which are then merged together. The space between the slices acquired can be, and usually is, different to the physical space between voxels in the slice. To alleviate issues associated with this, algorithms usually transform the images to have *isotropic* spacing, i.e., to have the same spacing in all dimensions, and subsequently transform the results back to the original space after the computation. The process used to transform the size and spacing of an image is called "*resampling*".

2.3. Medical imaging technologies

2.3.1. NIfTI

NIfTI-1 and NIfTI-2 (.nii or .nii.gz) are data formats used for storing medical images. They are adopted in research because they store data in a straightforward way and the standard is strict. However, images stored this way lack a lot of meta-information which hinders their clinical adoption.

2.3.2. DICOM

DICOM (Digital Imaging and Communications in Medicine) [22] is a standard for storing medical images and meta-information about them. It is used in clinical environments and it is becoming increasingly adopted in research too. Most medical scanners produce their results in the DICOM format. Meta-information can be stored inside the header in key/value (called tag/attribute) pairs. For instance, image modality is stored in tag "(0008,0060)". Other examples of information that can be stored include the name of the patient, date of acquisition, study in which the image belongs to, and acquisition protocol. Images stored in the DICOM format can be a single file, or multiple files where each contains only one of the slices. The latter was more useful in the past since it enabled sharing a large image using multiple storage devices, such as DVDs. It should be noted that a lot of the information in the header falls under PHI (Protected Health Information) and should be edited to be anonymized/randomized when used for research purposes. DICOM however has received criticism because it is not a strict standard and different companies might store data in a different way. As a result, a lot of applications, like MITK [23], have a set of different DICOM readers, from different organizations, and use the most compatible one automatically.

2.3.3. PACS

PACS (Picture Archiving and Communication System) is a technology used for storing and retrieving medical images across a hospital system that uses the DICOM format. It is composed of a central server where all the data are stored and various clients, i.e., computer programs, that run on the clinicians' computers. In that way, images don't have to be transported using physical devices across the hospital, but can be accessed instantly through the PACS client by anyone that has been granted access to them. A PACS also has the ability to display medical images directly, without the need of external software. Furthermore, with the help of remote access technologies, such as VPNs, a PACS can allow secure off-site access to the images for the clinicians. Security of the network and the

infrastructure is highly important because private patient information can be transmitted and accessed. Additionally, given that it is a highly automated and centralized system, a PACS can greatly reduce administrative costs.

In recent years, there has been an increase in PACS servers hosted in the cloud by dedicated companies. These companies also take care of the client software that runs on the practitioners workstation. By using a cloud based PACS, the hospital offloads the maintenance of the system to a company in exchange for a periodic fee. This move could also increase the security of the data, because a company that specializes in developing PACS has a dedicated security department that a hospital could not afford.

2.4. Clinical background

In this thesis, four cohorts were analyzed, consisting of glioblastomas, lung nodules, breast tumors, and images of spleen.

2.4.1. Cancer

Cancer is a group of diseases involving uncontrolled growth of cells in the body. According to the Global Cancer Observatory of the International Agency for Research on Cancer, there were over 18 million new cancer diagnoses in 2018 worldwide, while over 9 million people lost their lives due to the disease in the same year [24].

The main causes of cancer are tobacco use which accounts for 25-30% of cancer related deaths, diet which accounts for 30-35%, infections which account for 15-20%, genetics which account for 5-10%, and the remaining cases are attributed to other factors such as radiation, stress, physical activity, and environmental pollutants [25]. Since lifestyle is a main contributing factor, a lot of deaths could be prevented by quitting smoking, improving diet, vaccination, exercising, and avoiding direct exposure to sunlight. There are hundreds of types of cancer, but they can get grouped into the 5 main categories [26] described below.

Carcinoma starts in epithelial cells, which can be found on the skin or in thin layers in blood vessels or the exterior of organs. They also cover cavities, such as the chest and abdominal cavities. Most carcinomas form solid tumors [27]. Lung, breast, and prostate cancer are carcinomas. It is the most common type of cancer, accounting for about 85% of cancers in the UK [26].

Sarcoma starts in connective tissue, such as bones, nerves, muscles that support organs [27]. It accounts for about 1% of cases [26]. It can be further divided into bone sarcoma and soft tissue sarcoma.

Leukaemia is the cancer of white blood cells. It is caused by the bone marrow, which creates malfunctioning white blood cells in excess, causing them to build up in the bloodstream. Leukaemia accounts for about 3% of cases, but is the most common cancer in children [26].

Lymphoma is a cancer that begins in cells of the immune system, particularly in lymphatic cells. The lymphatic system is a network of vessels and glands which filters body fluids and fights infection; it consists of the lymph glands, the lymphatic vessels, and the spleen. The malfunctioning lymphocytes that are rapidly dividing can build up in the bone marrow and spleen and cause tumors. Lymphomas account for about 5% of cases. Myelomas start in plasma cells, which are a type of white blood cells used to produce antibodies and cause them to multiply rapidly. They account for about 1% of cases [26].

Brain and spinal cord cancers start in cells of the central nervous system. Glioma, the most common type, is a brain tumor that develops from glial cells. Brain tumors are particularly dangerous because they can impact parts of the brain that are important for life and early detection is hard. These cancers account for about 3% of cases [26].

2.4.2. Glioblastoma

Glioblastomas are a class of malignant brain tumors. They are classified as “Grade IV” on the WHO (World Health Organization) classification scale [28]. To be classified as “Grade

IV" a tumor type has to show anaplasia, mitotic activity with microvascular proliferation, and/or necrosis. Glioblastomas are known to reproduce rapidly and expand aggressively [28], [29]. According to a study in England that examined 10,743 patients with glioblastoma [30], median overall survival was 6.1 months, and the patients had 28.4%, 11.5% and 3.4% chance to survive 1, 2, and 5 years respectively.

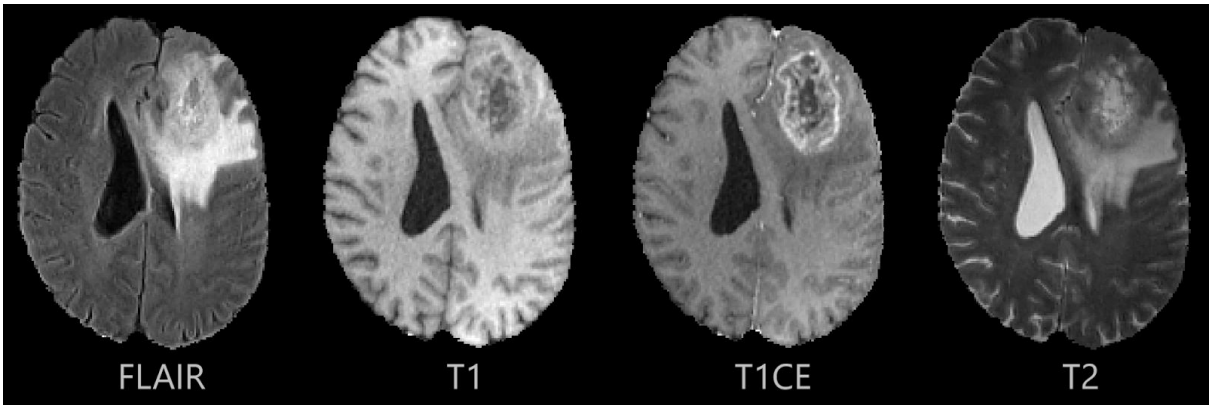

Figure 1: Glioblastoma images, as seen through different MRI acquisition protocols. ET appears bright in T1CE, NE is the dark area enclosed by ET in T1CE. ED is all voxels that are bright in FLAIR, that are not part of NE and ET.

Glioblastomas are large and have a complex structure. They are divided into two main sections, the tumor core (TC) and the peritumoral edematous/invaded tissue (ED). The tumor core is also divided into the non-enhancing (NE) and the enhancing (ET) parts [31], [32]. The tumor core subregions get their names from whether or not they show high intensity in the T1CE modality, after the administration of a contrast agent, however there are also underlying differences. The non-enhancing part consists of mostly pre-necrotic and necrotic regions of the tumor core, while the enhancing part is considered to be more active [32]. Although, the non-enhancing part is also considered to be infiltrative to some extent [33]. Lastly, the peritumoral edematous/invaded tissue surrounds the tumor core and has a hyper-intense signal in T2-FLAIR. It describes the regions of edematous white matter into the subcortex of the gyri. The ED region can also be infiltrated. The amount of infiltration can

be variable throughout the ED region and depends on the distance of an area to the tumor core [6]. As a result, sections of the peritumoral edematous tissue also get removed as a precaution, in case of tumor core removal surgery.

2.4.3. Lung nodules

Lung nodules are small masses of tissue in the lungs, which can be either cancerous or benign. They are typically less than 30mm in diameter and are visible in either CT scans or X-Rays [34], [35]. It is difficult to detect lung nodules and classify them as cancerous or not [36], because of their size and the similarity between malignant and benign nodules. Also, due to their small size, errors and inaccuracies in their segmentation heavily impact further computational analysis [37], as it can impact volume calculation and feature extraction.

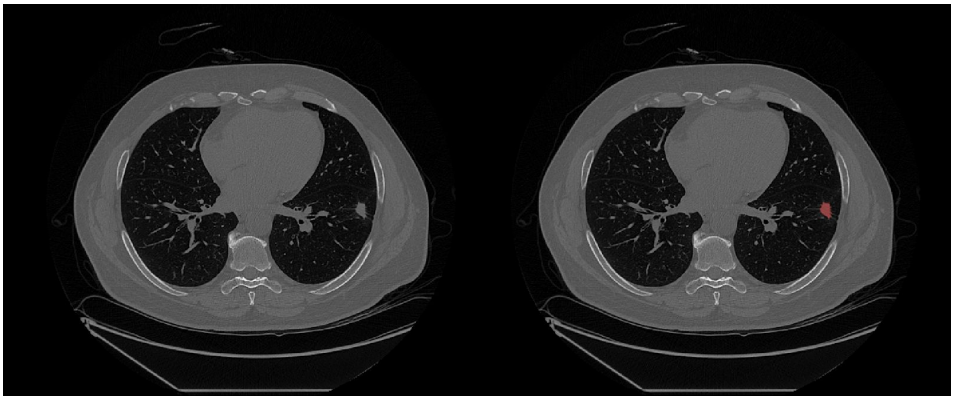

Figure 2: Lung nodule present in a CT image. The nodule is highlighted with red on the right.

2.4.4. Breast cancer

Breast cancer is a disease that affects a large percentage of the female population. In the US, approximately 41,760 women died from it in 2019 and around 268,600 women had an invasive case [38]. Prevention and early detection are very important in the fight against the disease. In this thesis, MRI images of malignant breast tumor cases were analyzed. There

were three images used for each case, one acquired before the administration of contrast agent and two after.

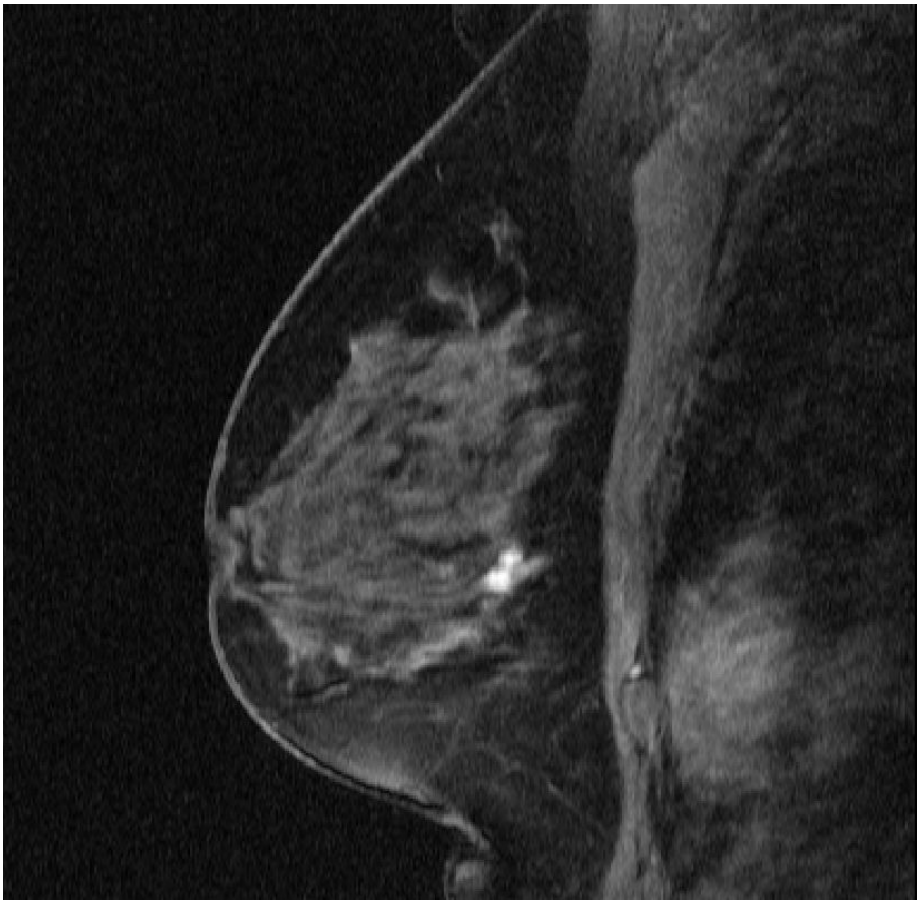

Figure 3: The small and bright area at the center of the mammogram is a breast tumor.

2.4.5. Spleen segmentation

Spleen segmentation was performed as an example of organ segmentation, which has a lot of applications. Specifically for spleen, its volume as well as changes in its size, can impact clinical decisions [39], [40]. It should also be noted that organ segmentation differs significantly from the segmentation of tumors, because organs tend to be larger with smoother boundaries.

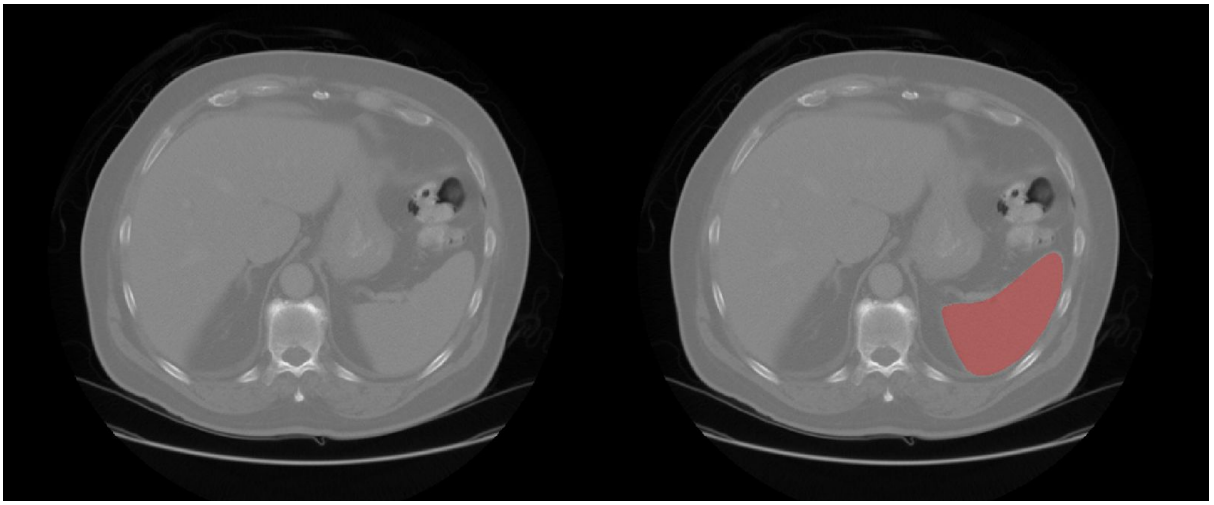

Figure 4: Image of spleen in CT (highlighted area).

2.5. Support Vector Machines

The method described in this thesis uses an ensemble [41] of Support Vector Machines (SVMs) [42] to make predictions. SVMs are used to perform multi-class supervised classification of individual pixels/voxels. However, SVMs can also be used for regression in other tasks.

The underlying theory of SVMs is best understood for 2-class classification. Each sample is mapped as a point in a n-dimensional space, where n is the number of features. SVMs try to separate the samples using a (n-1)-dimensional hyperplane, typically by choosing a hyperplane that best separates the training data into 2 classes. During inference, samples are mapped in the same n-dimensional space and samples that reside in opposite sides of the hyperplane get assigned to different classes. As an example, if there are two features, the separating hyperplane would be a line.

Linear separation is fast but not powerful enough to provide optimal results for most problems. Such separation happens in the example above, where a line is used to separate samples with two features. There is however a need for using curves instead of lines, i.e., non-linear separation. To address this, SVMs utilize the “kernel trick”. Using non-linear

kernel functions that are not excessively computationally heavy, the feature space can be transformed to another, higher-dimensional, space where the features are transformed and combined using the kernel function. As a result, the linear way of separation described in the previous paragraph can be used on the new feature space to perform non-linear separation. The choice and parameterization of the kernel function is application-specific.

2.6. Overview of segmentation methods

A supervised ML model trains on labeled input data and produces a prediction (output) on new unlabeled data. Unsupervised learning uses no labeled input data, but rather tries to find patterns in the data and group them accordingly. Semi-supervised approaches use a few labeled input data alongside a large amount of unlabeled ones, which is useful when there is a large amount of available data, but few of them are manually annotated and annotating them would be costly. Most state-of-the-art fully automatic methods are based on deep learning. These methods typically need large annotated datasets.

An algorithm that has traditionally been used for supervised segmentation is k-nearest neighbors (kNN). It operates on the assumption that similar objects tend to be close in a multidimensional feature space. It is a slow algorithm that requires the user to select k, through experimentation. At classification, the k nearest (to the test sample) training samples are selected and the most often represented class in the k near training samples is selected. Selecting an odd k is recommended to avoid ties.

Another similar algorithm that has been used are SVMs, which were analyzed in a separate section. SVMs could also be adapted to perform semi-supervised classification, such as the Transductive Support Vector Machine (TSVM) [43], where the algorithm can utilize both labeled training data and unlabeled ones to improve the model.

Ensemble methods combine multiple models to achieve better results than what each one of them would have produced individually. Usually, ensemble methods combine weak learners, i.e., methods which typically have high bias or variance, to create a strong learner. Because of the bias-variance tradeoff, individual models with low bias have high variance, and vice versa, but through combining some the effect is minimized. There are three main ways to combine the results of the multiple classifiers: bagging, boosting, and stacking. In bagging, the weak classifiers are trained independently and the final result is the average of all classifiers. To train each classifier, random samples of the training data are chosen with replacement, meaning that samples chosen for one classifier have the same probability to be chosen for another. Random forests are a popular bagging technique. Bagging reduces variance. In boosting, samples are weighted, meaning that there is a variable probability for them to be chosen to train a classifier. That is because training in boosting is a sequential procedure and not a parallel one like bagging. When a classifier is trained in boosting, the samples that were misclassified get assigned a higher weight. Thus, classifiers trained afterwards have a higher chance to be trained on the previously misclassified samples, which reduces bias. Stacking works with heterogeneous weak learners, i.e., combining different algorithms. In that case a new model (meta-learner) is trained that accepts the results of the different weak learners as input.

Perceptron is an algorithm for binary supervised classification. It accepts multiple inputs, which are subsequently individually weighted. The weighted inputs are summed and passed through an activation function. A bias could also be added to help shift the outcome of the activation function. The activation function determines if the output is classified as 1 or 0. There are two types of activation functions, linear and nonlinear. Usually non-linear activation functions are selected. Some popular ones comprise the Sigmoid, Tanh, and ReLU (Rectified Linear Unit) functions, with the latter being popular in deep learning.

Training a perceptron is an iterative procedure. It is an optimization problem, where the weights of the perceptron are updated to minimize an error function.

Artificial Neural Networks (ANN) are structures with three layers. The first layer is the input, the last one is the output, and the middle one is called a hidden layer. Neurons (nodes) in the hidden and output layer are perceptrons, while nodes in the input layer contain the numerical value for each feature of the sample. All neurons in the hidden layer have weighted input from all the input neurons. Additionally, all neurons in the hidden layer are connected to all output neurons. Since multiple output neurons can exist, multi-label classification is now possible, which was not the case for the perceptron. The weights of the neurons are updated in training. An error value is calculated through a cost function, according to the output value produced and the expected value. Through back-propagation, weights are adjusted to minimize the cost function.

Deep neural networks are artificial neural networks that have two or more hidden layers. Deep networks typically don't require any feature extraction beforehand and can learn the patterns of the data by essentially generating their own features. Their adoption has exploded in the last two decades, because of advances in computing power, data availability and ML techniques.

Convolutional Neural Networks (CNNs) are useful in image recognition tasks, because they can accept a whole image as input, which for medical imaging can account to millions of pixels, but manage to keep the complexity of the network from getting high. Usually a CNN has three layers: a convolutional, a pooling, and a fully-connected layer. In the convolutional layer, the dot product of a filter and the matrix of an image portion is calculated, with the aim of extracting features. The pooling layer decreases the dimensionality of the feature matrix, by combining squares of the feature matrix, usually by keeping their maximum (max-pooling) or their average (average-pooling). After these two layers, the feature matrix has been

reduced significantly and a fully-connected layer is used to learn the relationships of the data, like the hidden layer in an artificial neural network.

The convolution neural network described above is useful for classification, but not segmentation. To adapt it to segmentation, the U-Net and other similar architectures have been proposed and used successfully in medical imaging. The difference is that apart from the contracting path that reduces the size of the data, there is also an expansive path that increases it and brings it back to the original image size (excluding channels). That is because in segmentation a label needs to be assigned to each individual pixel/voxel.

Transfer learning is a machine learning technique used primarily for domain adaptation. It is more useful in MRI, because there is more variation in the image intensities and contrast [44]. Thus, models don't perform well on images acquired with different machines/protocols than the one they were trained on. The main idea of transfer learning is to use large legacy datasets for training models and then use a small number of annotated scans from the target domain to adapt the model. Adaptive SVM and TrAdaBoost (Transfer AdaBoost) have been shown to outperform traditional SVM classifiers in medical image segmentation [45] even when trained on a small number of images from the target domain. Domain adaptation has been shown to work well with CNNs too [44]. Transfer learning could also work with models originally trained for non-medical tasks, then adapted to medical segmentation [46].

BraTS (Brain Tumor Segmentation) [32] is a yearly multi-institute effort to acquire medical images of brain tumors and use them to assess the robustness of segmentation algorithms in the form of a challenge. Similar challenges exist for other segmentation tasks too [47]–[49]. Automated methods used in BraTS have great performance, but are still less robust than segmentations obtained manually from clinicians, with less accuracy and inter-rater agreement. However, fusing the results of multiple algorithms, through label

voting, can obtain better results than the individual algorithms and inter-rater agreement that is comparable to the experts.

2.7. CAD Systems

A software system can't replace the radiologist in most situations yet. This would require exceptional sensitivity, specificity, and adaptability from the algorithm. Additionally, the general population might not be ready for such a fundamental change and insurance companies might be hesitant to cover such treatments. On the other hand, physicians are subject to mistakes too. Such mistakes could be triggered by fatigue, stress, and other factors [50], [51]. Thus, a physician could benefit from a software solution that is not perfect but correct enough in most situations, as the algorithmic decision could act as a suggestion or an alternative opinion, assisting the radiologist but leaving the final decision to them [52]–[55]. Such a system is called a CAD, which is an umbrella term that encapsulates computer-aided detection (CADe) and computer-aided diagnosis (CADx) [56]. CADe systems aim to detect unusual structures in the body, such as tumors. CADx systems aim to make intelligent decisions about the patient's health, such as classifying tumors as malignant or not. While the two types can exist separately, a lot of CAD systems encapsulate both CADe and CADx components.

While the first attempts in using computers to solve medical problems occured in the 1950s, by the 1970s it had become clear that it is not an easy problem to solve using traditional computer techniques and statistics [56]. In more recent times, advances in computer power and machine learning techniques have made the exploration of the field plausible again and have spawned a lot of research on the subject.

2.7.1. Data collection

If a CAD system requires training, as most intelligent ones do, a large dataset is required. This dataset should preferably contain images taken from multiple institutions, various machines, and different cohorts. This data needs to be checked to exclude scans with missing images and outliers. Alternatively, missing scans can be populated with data augmentation techniques [57]. Training data need to be normalized and potentially harmonized to assist the model in learning the actual differences.

2.7.2. Step 1: Preprocessing

In the first step of a CAD system, data should be normalized/standardized and potentially harmonized if they have been acquired using different scanners/techniques. Normalization pertains to scaling the data in the [0,1] range and standardization to transforming the data to have zero mean and a standard deviation of one. Data harmonization pertains to minimizing differences between images acquired from different scanners [58] so that the model can learn the real differences between cases. Additional processing can occur in this step depending on the application, that aims to make the data easier to be processed later, such as denoising techniques. For brain applications there might be an additional *skull-stripping* step, where the skull is removed from the image, leaving only the brain to be processed.

2.7.3. Step 2: Segmentation

Segmentation of the images is needed, in order to isolate the structures that need to be analyzed. The segmentation algorithm can be automatic or interactive, although automatic methods are preferable if available for the specific problem, since no other CAD step typically requires human interaction.

2.7.4. Step 3: Feature extraction

Feature extraction typically follows segmentation, a process that aims to reduce the input data into values (features) that are relevant to the problem at hand [52], [54], [59], [60].

A popular method for feature extraction of medical images is called radiomics. "pyradiomics" [61], a popular radiomics python package, categorizes the features it can extract into the following categories: (i) "First Order" features which are related to the distribution of pixel/voxel intensities in the segmented area (ii) "Shape" features which describe the shape of the segmented area in 2D/3D (iii) "Gray Level Co-occurrence Matrix" (GLCM) features which are related to the second-order joint probability function of the segmented area (iv) "Gray Level Size Zone Matrix" (GLSZM) features which calculates the number of of gray level zones in the image, i.e., areas that have the same intensity values (v) "Gray Level Run Length Matrix" (GLRLM) features which calculate the length of consecutive pixels/voxels with the same intensity value (vi) "Neighbouring Gray Tone Difference Matrix" (NGTDM) features which calculate the difference between the intensity of a pixel/voxel and the average intensity value of its neighbours within a predefined distance and (vii) "Gray Level Dependence Matrix" features which calculate the number of pixels/voxels that are "dependent" on a pixel/voxel; a pixel/voxel is dependent on another if the absolute of their intensity difference is less than a predefined value.

One other popular feature extraction technique is Principal Component Analysis (PCA). In this context, PCA is used for dimensionality reduction of the information of the segmented area. A NxN symmetric covariance matrix is constructed, where N is the number of pixels/voxels, because a value is needed for each pixel/voxel pair. Each value in the matrix represents the covariances of the respective pair. If a value is positive then there is a positive correlation and the two variables increase/decrease together, while the reverse is true for negative values. PCA constructs new variables, called principal components, that are not correlated [62] and are the same in number as the input variables. However, PCA encloses as much information as it can in the first principal components, i.e., maximizes variance in the first components. As a result, by selecting a small number of the first principal components as features, most information can be kept while the dimensionality of the

problem reduces. Independent component analysis (ICA) is a similar technique used to separate a signal into underlying independent components. It doesn't focus on maximizing variance in the first components, but focuses on making the components independent. Another technique is the autoencoder, which is an unsupervised way to achieve dimensionality reduction by using an artificial neural network (ANN). By having a "bottleneck" in the hidden layer the information of the input image is compressed, because the network learns correlations of the input.

2.7.5. Step 4: Feature selection/elimination

"Dimensionality curse" is a term that describes issues pertaining to having data arranged in high-dimensional spaces. The main reasoning is that high-dimensional data tend to be sparse and that negatively impacts algorithms that need statistical significance. CAD systems can face such problems and as a result feature selection usually follows feature extraction [53], [57]. The most important features are selected from the feature space and they are used to train the model. With feature selection, training time gets lower too and there is less chance of overfitting. There are three main feature selection technique categories that are described below [63].

The first category is called a filter and these types of techniques don't interact with the classifier, that simply accepts the selected features as input. Filters use application-independent statistical methods to categorize the features. They can be further subdivided in univariate filters, that don't use feature dependencies in the calculations which can result in selecting similar features, and multivariate filters, that take notice of feature dependencies at the cost of processing speed. The most common filters are multivariate Pearson correlation-based.

Another category is embedded feature selection. These types of methods are specific to certain machine learning techniques and select features during training of the model. They

can be thought of as part of the classifier. An example of embedded feature selection is Recursive Feature Elimination for Support Vector Machines.

The last category is wrapper methods. These methods use the performance of the classifier to evaluate the suitability of the features. This is an iterative procedure that tends to be slower and less scalable than the other two, but can produce better results with machine learning techniques where embedded feature selection does not exist. An example is backward/forward elimination.

2.7.6. Step 5: Analysis

The final step is usually the main algorithm of the CAD that performs a classification or regression task. These algorithms have an output that can be interpreted by the clinician and usually is a decision about the health of the patient. It should be noted that some deep learning algorithms might be trained using the whole image as input and not require separate segmentation and feature extraction/selection.

2.7.7. Workflow of a CAD in a hospital environment

During usage in a hospital setting, a CAD system takes the medical images of a patient as input, potentially alongside additional meta-information such as the age of the patient. The first step is preprocessing the images to normalize/harmonize them with the training data. Segmentation, feature extraction/selection and analysis happen in accordance to the trained models and the result is presented to the clinicians. Excluding the case where the segmentation step is interactive, this process can be automated and transparent to the clinician, who does not need to be aware of the process. In modern systems, CAD software can run directly in the PACS, enabling the clinician to have the segmentation and the generated diagnosis/detection available for him when he sees the scans for the first time.

2.8. Adaptive geodesic distance maps

Adaptive geodesic distance (AGD) [16], [64] maps are utilized in the proposed method as additional input to the model, alongside medical images. Their purpose is to convey additional information related to the structure that a model, which uses pixels/voxels as training samples, would struggle to deduce.

To produce an AGD map, two images, having exactly the same size, are needed: (i) a medical image, (ii) a binary image, i.e., an image with values 0 or 1 at each pixel/voxel. Since these images have the same size, each pixel/voxel in the binary image can be directly mapped to a particular pixel/voxel in the medical image. The positive values in the binary image represent the drawings of a human operator. More precisely, the positive pixels/voxels in the binary image represent samples of pixels/voxels in the medical image that the operator deemed as belonging to the structure that needs to be segmented.

Geodesics generalize the concept of straight lines to curved spaces. An example of geodesics would be trying to calculate the distance of two cities on earth. Their euclidean distance would be calculated as the length of a straight line connecting the two cities that passes through earth, while their geodesic distance would follow the curvature of the earth. The problem can be thought of as trying to find the shortest path between two points while moving along a surface.

AGD maps have a “distance” value at each pixel/voxel. This “distance” represents the shortest path between the relevant voxel and the closest pixel/voxel marked as positive by the operator. This distance is not euclidean; similar to geodesics it represents the shortest path between two points while moving along a surface. However, the surface is defined by the intensity profile of the medical image [16]. As a result, the distance values are a composite of physical distance and intensity change. In other words, pixels/voxels far away from the drawings and/or with different intensity have higher values.

The maps can be calculated rather quickly, by visiting each pixel/voxel twice [16]. The result is an image that has values in [0,+∞). Lower values represent a higher probability that the pixel/voxel belongs to the same structure as the ones that the operator drew upon. AGD maps have been used in the past to perform single-class interactive segmentation, which required the operator to threshold the map to produce the final segmentation. In detail, by choosing a value in [0,+∞), the values lower or equal than the chosen threshold would be transformed to 1 in the output segmentation, while all the other values to 0. This method was quick and worked well for simple structures, however it faced challenges on more complex applications, as it couldn't handle multiparametric images or perform multi-class segmentation. The thresholding step was also difficult for operators, as the correct value selection was not always obvious. This thesis incorporates AGD maps as additional input to a machine learning model, thus solving all the aforementioned shortcomings.

2.9. Statistical analysis concepts

Dice coefficient [65] quantifies the spatial overlap of two binary segmentations, by identifying the pixels/voxels that have a similar value in both segmentations and the ones that do not. When a segmentation is not binary, but has multiple labels representing different sub-regions of a single structure, the individual sub-regions can be converted to binary segmentations and compared. Similarly, all labels can also be merged into one to quantify overlap for the whole structure. For the dice coefficient calculation, the following formula was used for each sub-region as well as the whole structure. TP is the number of true positive pixels/voxels, FP are false positive, and FN are false negative.

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

The dice coefficient results did not follow a Gaussian distribution. The samples were also paired, because for each patient there were segmentations to compare from different operators, cycles, or time points. As a result, the non-parametric signed-rank paired Wilcoxon test was used to evaluate significance.

Finally, Pearson's Correlation Coefficient [66] was used to quantify the correlation between paired sets of *volumes* of segmentations acquired with different means.

3. Implementation

3.1. Cancer Imaging Phenomics Toolkit (CaPTk)

CaPTk [67] is an open-source cross-platform software tool focusing on cancer research, developed by the *Center for Biomedical Image Computing and Analytics (CBICA)* of the *University of Pennsylvania*. It provides functionality for preprocessing, segmentation, feature extraction, and training of machine learning models, as well as advanced specialized applications of computational studies conducted in the center. The implementation of the method is part of current releases of CaPTk and can be obtained through this link: https://www.cbica.upenn.edu/captk.

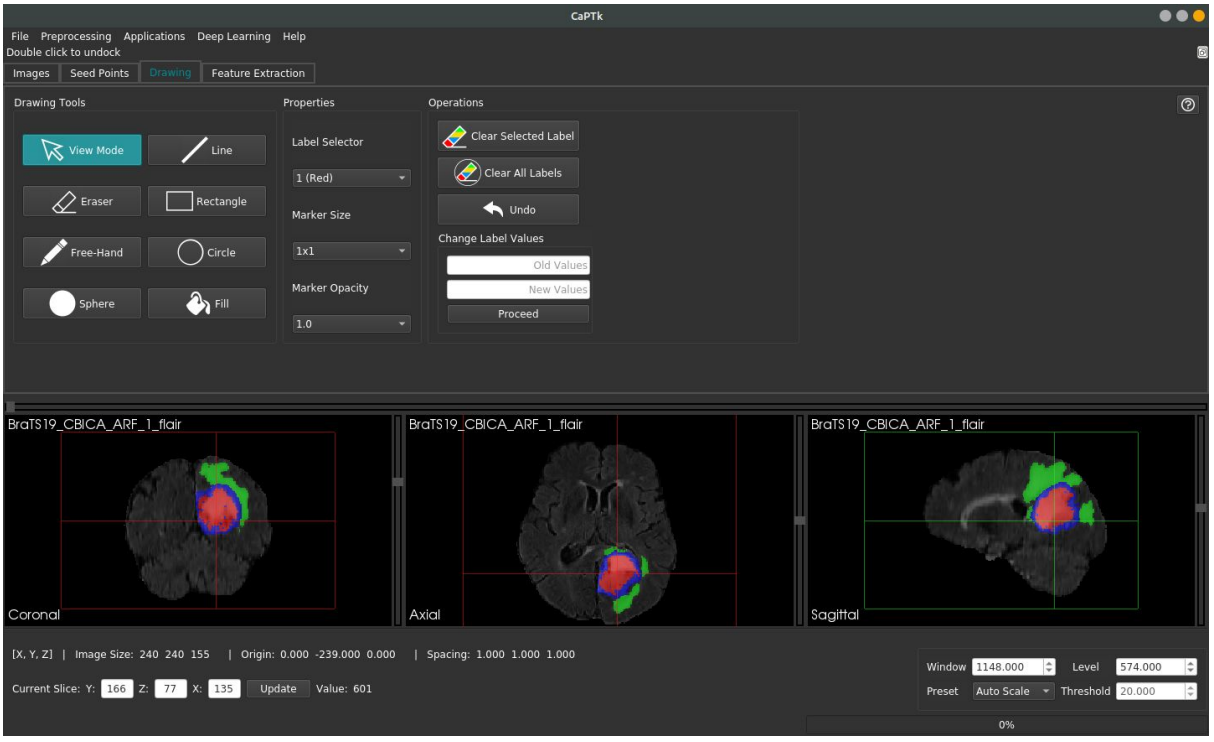

Figure 5: The graphical interface of CaPTk.

3.2. MITK Plugin

MITK [23] is an open-source cross-platform application framework for medical image processing, developed by the *German Cancer Research Center (DKFZ)*, that has seen success in both research and industrial environments. MITK supports extensions via a plugin mechanism, which was used to implement the algorithm. An MITK application (*MITK Workbench*) that contains the developed method as a plugin can be found at: https://github.com/CBICA/InteractiveSegmentation.

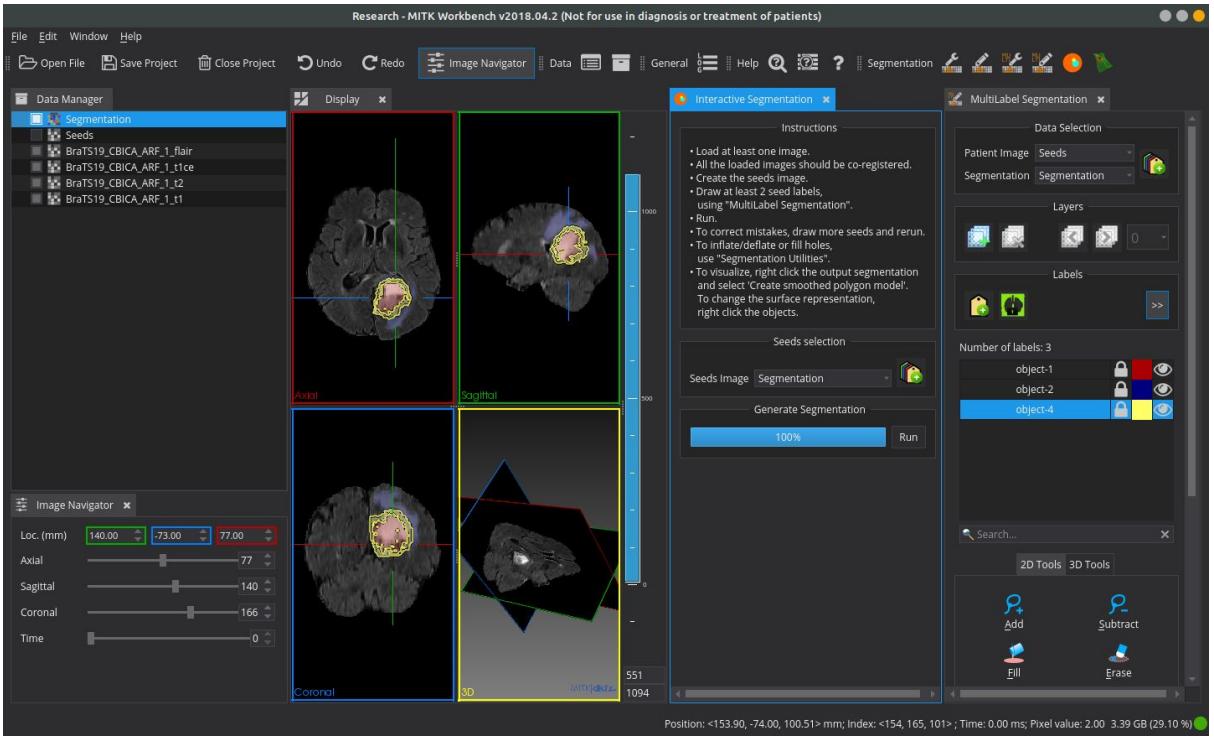

Figure 6: The graphical interface of *MITK Workbench*, featuring the developed plugin.

3.3. Software libraries used in the implementation

Both algorithm implementations are in C++ and CMake was used to control the software compilation process. Additional software libraries that were used are detailed below.

3.3.1. ITK

ITK [68], [69] is an open-source software library that provides tools for medical image analysis. It provides functionality for loading and saving n-dimensional *itk::Image(s)*, iterating over their pixels/voxels, and performing multiple operations which are implemented as "ITK filters". Such relevant filters are resampling, normalization, standardization, and image denoising. ITK version 4.0+ can be used with the source code of the algorithm.

3.3.2. OpenCV

OpenCV [70] is an open-source software library for computer vision and machine learning. The class "cv::ml::SVM" was used to implement the SVMs used in the implementation. To provide input to OpenCV, the images are transformed to *cv::Mat*, OpenCV's matrix implementation, and the results are transformed back to *ITK images* after the computation. The three kernels used comprised *RBF* (Radial Basis Function), *CHI2* (exponential χ^2), and *INTER* (Histogram Intersection). The *trainAuto()* function was used to perform cross-validation, hyperparameter tuning, and training. The default parameter grids of OpenCV were used for the grid search performed in each cycle of cross-validation, except for the C parameter of *RBF* and *CHI2* where the values tested ranged [1,400] with a logarithmic step of 2.5.

3.3.3. MITK

MITK [23] was used to develop the MITK plugin of the algorithm. It is a plugin-based application framework that contains various libraries commonly used in medical imaging tasks, such as ITK and OpenCV. MITK also provides additional high-level functionality, such as better image I/O functions and image data containers, as well as various graphical components like an image viewer and a data manager.

3.4. Software libraries for statistical analysis

The experiments required a complex statistical analysis that would have been difficult to perform without a dedicated codebase, given that multiple calculations were needed for each cohort, rater, segmentation cycle, and structure sub-region.

Python 3.6 was used to perform the statistical analysis. The following packages were used: (1) *pandas 1.0.3* [71] for storing/retrieving dice coefficients in CSV form and performing high-level operations on them. (2) *numpy 1.18.4* [72] for performing a vast range of array operations using the data. Correlation coefficient was also calculated using numpy through the corrcoef() function. (3) *scipy 1.4.1* [73] for non-parametric paired wilcoxon tests, as well as some operations of Dice calculation. (4) *matplotlib 3.1.3* [74] for creating the plots.

4. Materials and Methods

4.1. Data

Experiments were approved by the Institutional Review Board (IRB) of the University of Pennsylvania (UPenn, Philadelphia, USA). Quantitative evaluation was based on public and private clinical data from four retrospective cohorts (Spleen (3D-CT, n=20/41, Medical Segmentation Decathlon [75]); Breast tumor (2D-DCE-MRI, n=50, multimodality trial at UPenn; NIH P01CA85484); Lung nodules (2D-CT, n=50/89, The Cancer Imaging Archive [76]–[78]); Brain Glioblastoma (3D-MRI, n=20/335, BraTS'19 [3], [4], [32])). Cohort subsets were created, following random selection, to facilitate the exhaustive manual annotations described hereafter. The Brain (11 Males / 9 Females: Age=62.84/64.36, Range=44.82-77.48/39.64-77.09) and Breast (Female: Age=50.41, Range=32.68-71.97) datasets were acquired from 2006-2014 and 2002-2006, respectively. The Spleen (13 Males / 7 Females: Age=63.85/58, Range=40-81/48-68) and Lung (34 Males / 16 Female) were acquired from 2000-2013 and 2004-2011, respectively. Age information was not available for the Lung dataset. Ground truth segmentations were available for all datasets, except for Lung which were created by a fellowship trained, board certified, thoracic radiologist (S.K., 21 years of experience).

4.2. Proposed Segmentation Algorithm

The algorithm can segment **N** regions of interest (ROIs) at one time by initializing **N+1** different labels, where the additional one accounts for the "background". As a first step, the user briefly draws over the different ROIs using at least two distinct labels (Fig.7).

Table 1: Outline of the datasets used.

Cohort	**Modalities**	**Number of cases**
Spleen	3D-CT	20
Breast	2D-DCE-MRI: pre-contrast first post-contrast second post-contrast	50
Lung	2D-CT	50
Glioblastoma	3D-MRI: T1 T2 T1CE (T1Gd) FLAIR	20

For each pair of image and class labels, an AGD map [16] (Fig.8) is produced reflecting a composite of intensity and spatial distance from the drawings, such that voxels far away and/or with very different intensity have higher values (in the figure, such voxels appear brighter). Additionally, for the purpose of providing more spatial information, three "coordinate" maps are used, one for each dimension of the image, where the values range from 0 to the size of the image in that dimension.

An ensemble [41] of SVMs is trained on voxels that belong to the drawings and segments the remainder of the scan. Each training sample (i.e., voxel) is described by the following features: (i) intensity across all co-registered images, (ii) distance in all AGD maps, and (iii) value in all coordinate maps. Three SVM classification models (i.e., radial basis function (RBF), chi-squared, histogram intersection kernels) are trained and their hyperparameters are selected through cross-validation. Each voxel's final prediction is obtained by fusing the three model predictions via majority voting and the RBF classifier is used to resolve ties.

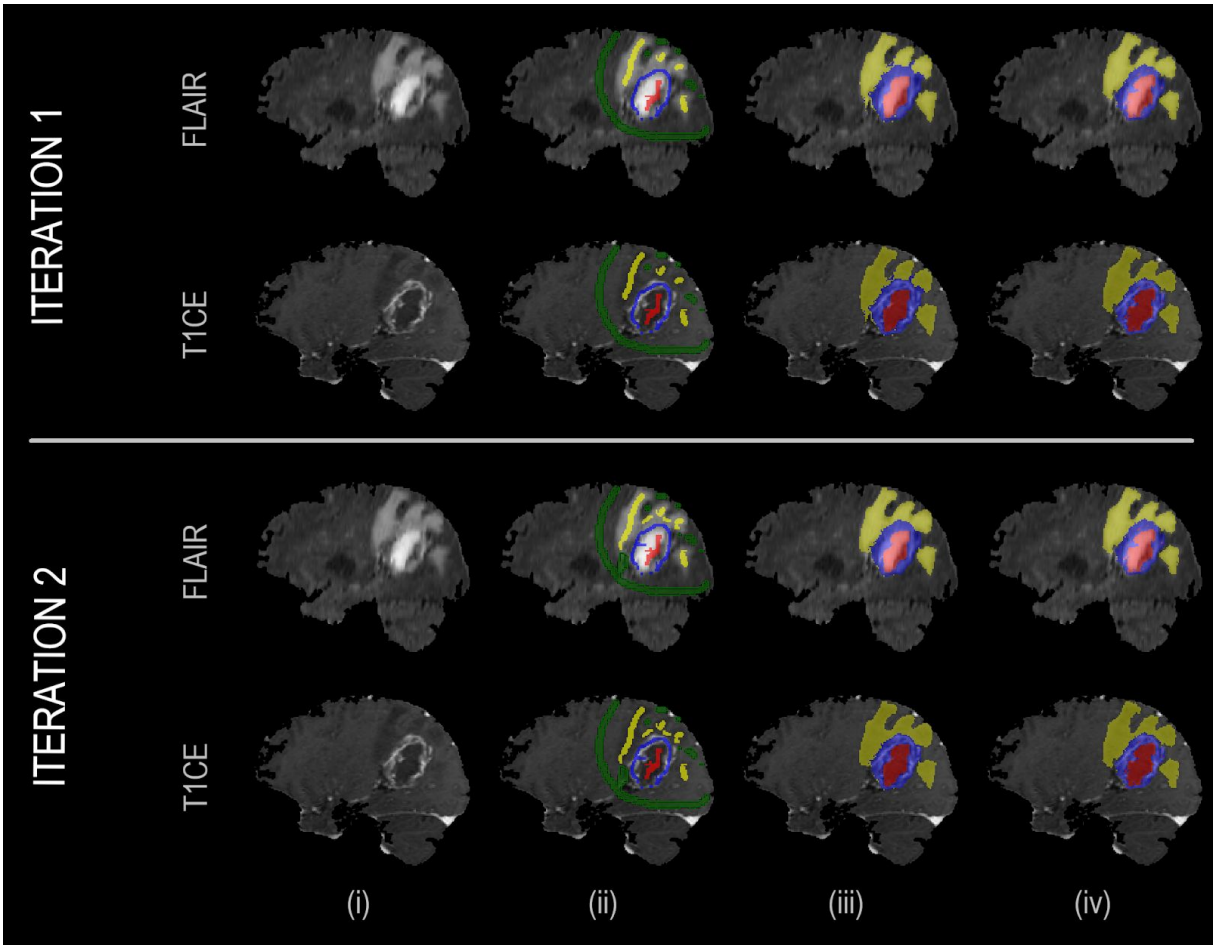

Figure 7: Example showcasing the result improving as a function of invested time. In the first iteration the user quickly draws over the different areas. In the second iteration, the user places few additional labels to correct representative misclassified areas, which are then used to retrain the machine learning model. From left to right: (i) Anatomical image (ii) User annotations (iii) result segmentation (iv) ground truth segmentation.

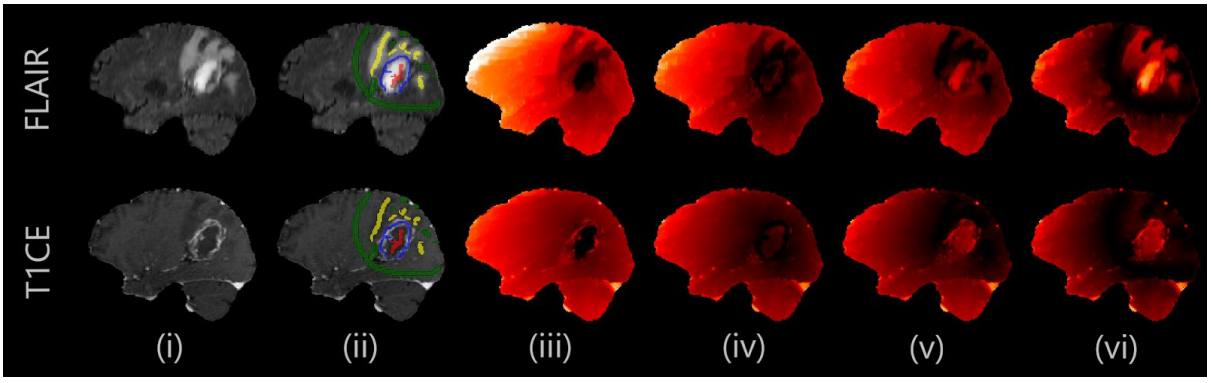

Figure 8: Example of AGD maps. Darker values indicate lower adaptive geodesic distance from the user drawings. In glioblastomas, NE and ET boundaries are clearer in T1CE, while the boundary between ED and background is clearer in FLAIR. From left to right: (i) Anatomical image (ii) User annotations (iii) AGD map for NE (Non-enhancing tumor core) annotation (iv) AGD map for ET (Enhancing tumor core) (v) AGD map for ED (peritumoral edema) (vi) AGD map for background.

4.3. Experimental Design

4.3.1. Evaluation protocol

To quantitatively evaluate the method, eight experts were included, two for each cohort. Each expert was asked to segment every scan four times, thereby producing two manual and two IML-assisted annotations, in addition to the extensively defined ground truth segmentations. The experts were given brief instructions for the method and were asked to note average time for their segmentations. To have a fair assessment of inter-rater consistency for glioblastoma segmentation, the experts were instructed to perform the manual segmentation of the various tumor sub-regions (enhancing tumor (ET), non-enhancing tumor (NE), and peritumoral edematous/infiltrated tissue (ED) [3], [4]) in 1 hour or less.

4.3.2. Experiment 1: Overall performance evaluation

Initially, the spatial overlap agreement of each approach was evaluated relative to the ground truth by utilizing the Dice Similarity Coefficient (DSC) as a metric to select one IML-assisted and one manual segmentation from each rater. For glioblastoma, only the whole tumor (WT) area was used for these selections. The DSCs of the two sets were statistically compared. Additionally, volumes calculated for IML and manual segmentations were quantitatively compared with the ground truth. Lastly, the Pearson's Correlation Coefficient [66] was estimated for each of the paired segmentations separately, i.e., (i) IML vs ground truth and (ii) manual vs ground truth. The average active drawing time was compared for each cohort between IML-assisted and manual segmentation.

4.3.3. Experiment 2: Intra-rater segmentation consistency

The DSCs within the two IML-assisted and the two manual segmentations of each rater were calculated (i.e., $DSC_{IML1/IML2}$, and $DSC_{Manual1/Manual2}$) for each case. The DSCs were statistically compared between the IML-assisted and manual segmentations. Furthermore, the existence

of significant differences between the DSC of the manual and IML-assisted segmentations relative to ground truth was also examined for each rater separately.

4.3.4. Experiment 3 - Inter-rater segmentation consistency

The best segmentations of each rater were selected with the same selection criteria as experiment 1. The DSCs within the best IML-assisted segmentations across raters, and within their best manual annotations, were calculated for each case, and their significant differences were evaluated.

4.3.5. Statistical Analysis

Paired Wilcoxon-signed rank non-parametric statistical tests [79] were used for statistical comparisons (assuming a type I error rate of 0.05), because the samples were paired and tended not to follow a Gaussian distribution.

5. Results

In this section, the results of the experimental validation are presented for 20 spleen, 50 breast tumor, 50 lung tumor, and 20 glioblastoma cases. For each patient, each expert produced two manual and two IML-assisted segmentations, which were additional to the extensively defined ground truth segmentations.

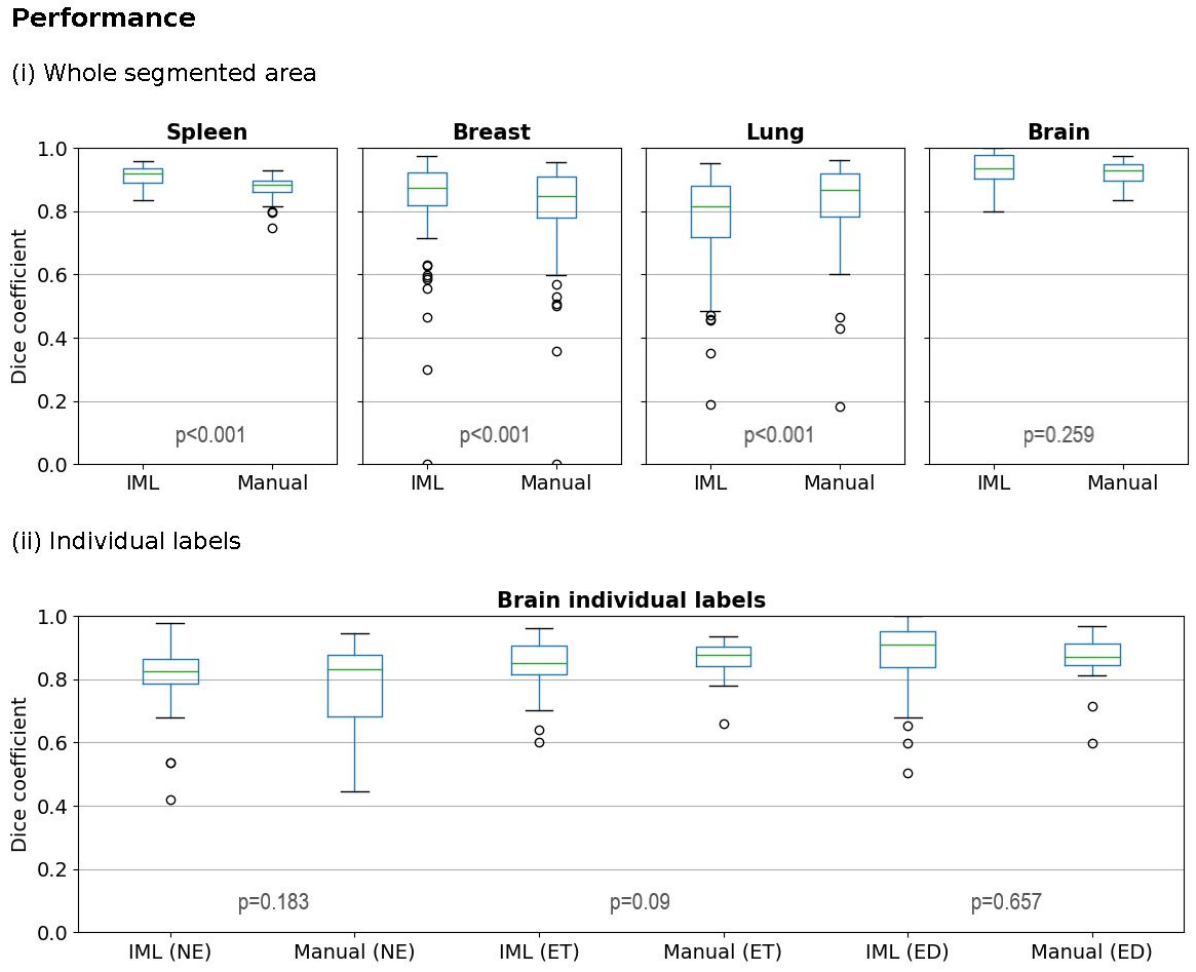

Figure 9: Dice coefficient, compared to ground truth, of: (i) all individual labels representing different areas of the structure counted as one, (ii) the individual areas of glioblastomas.

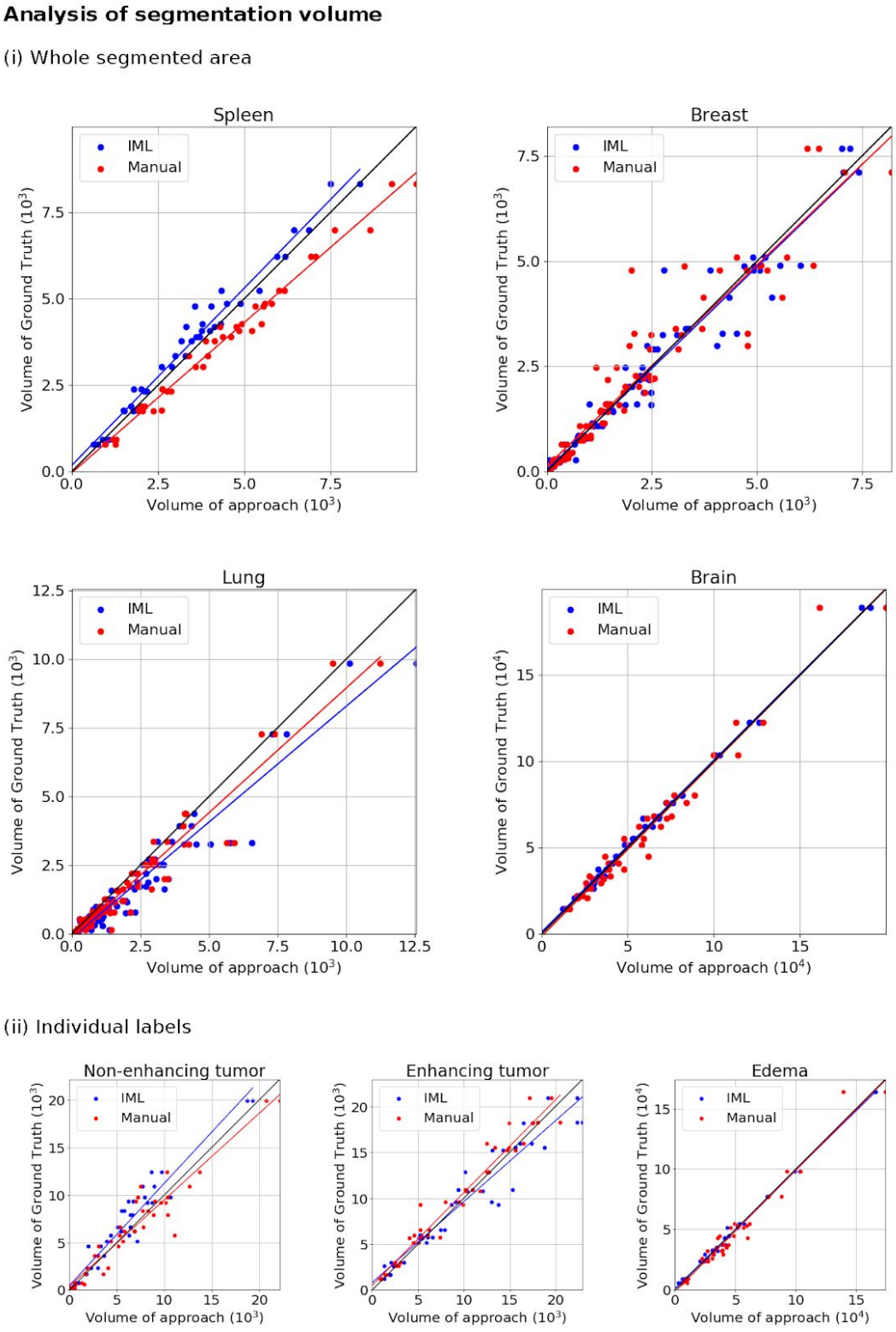

Figure 10: Scatterplots in which blue points are the pairs of volume of IML method and volume of ground truth and red are the pairs of manual segmentation volume and ground truth. The black line represents the ground truth's volume. The plots belong to (i) different cohorts where all individual labels, representing different areas of the structure, are counted as one, (ii) the sub-regions of glioblastomas.

5.1. Experiment 1: Overall Performance Evaluation

In the first experiment, the performance of the proposed method was evaluated (Table 3, Fig.9). For glioblastomas, manual and IML-assisted segmentations yielded similar pairs of DSCs both for WT and individual sub-regions, thereby indicating no significant difference between them, whereas the converse was true for other cohorts. The method achieved higher DSC on average for spleen and breast tumors, but lower for lung nodules when compared with the manual segmentations. However, the method was substantially faster than manual annotation in all cohorts (Table 3), by **53.1%** on average.

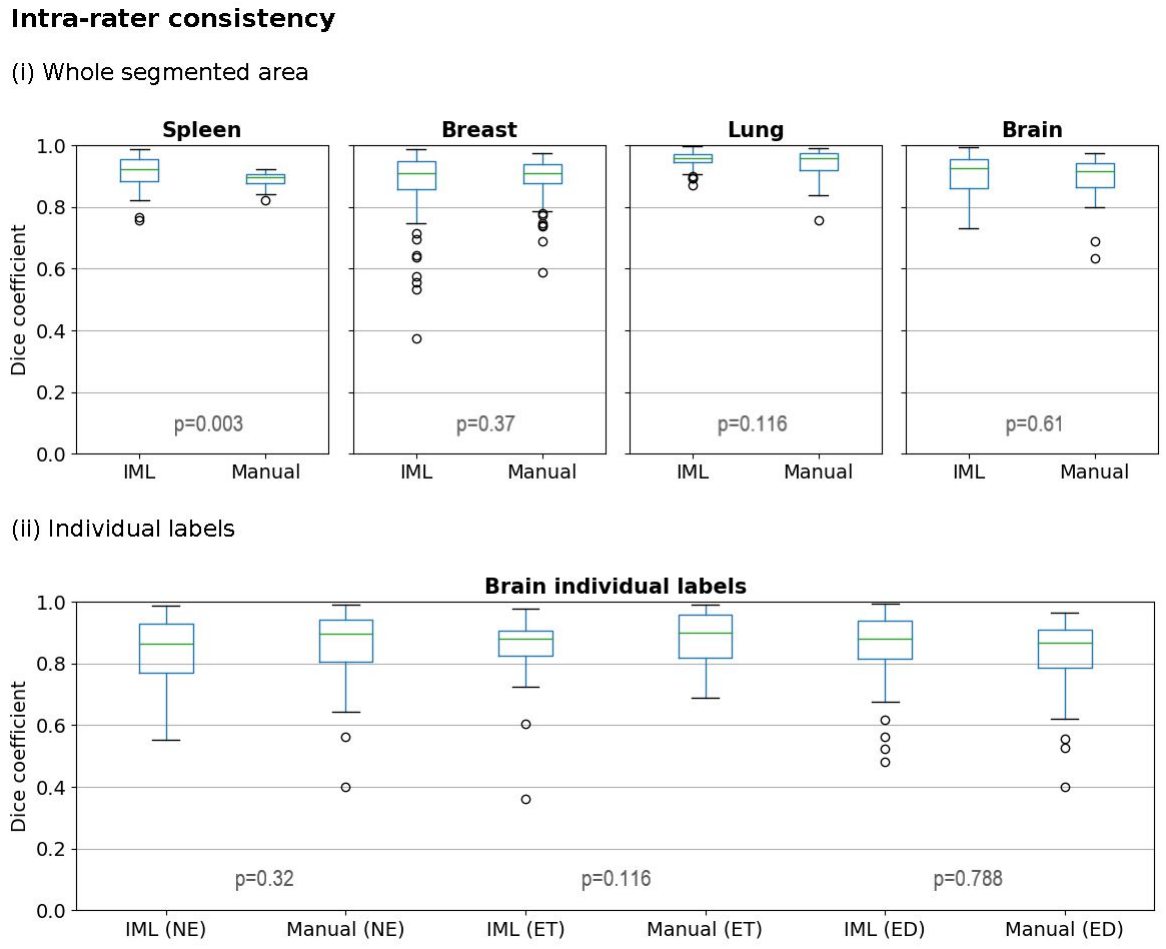

Figure 11: Dice coefficient, for intra-rater consistency, between the first and second round of the raters of: (i) all individual labels representing different areas of the structure counted as one, (ii) the individual areas of glioblastomas.

An analysis of the volume of ground truth, manual and IML-assisted segmentations (Fig.10, Table 3 “Corr. Coef.” column) shows that errors made by the method were mostly systematic. This is more evident in spleen images, where IML-assisted and manual segmentations revealed systematic under- and over-segmentation, respectively. The method made some non-systematic errors in lung nodules and the ET glioblastoma sub-region, but these areas were also more erroneous in manual segmentations. Notably, ET is regarded as the most challenging area of glioblastoma, because it frequently has unclear and smooth boundaries [3].

Table 2: Results (p-values) of a paired Wilcoxon test for each rater, comparing the dice coefficient results of the different approaches relative to ground. $p<0.05$ indicates a significant difference.

	Rater 1		**Rater 2**	
Label	IML	Manual	IML	Manual
	Spleen			
-	0.9405	0.3507	0.1454	0.433
	Breast			
-	0.2425	0.5116	0.5921	0.1358
	Lung			
-	0.0422	<0.0001	0.1358	<0.0001
	Brain			
WT	0.156	0.0001	0.8813	0.0008
NE	0.0522	0.0859	0.9405	0.0001
ET	0.1913	0.3317	0.0522	0.0001
ED	0.4781	0.0001	0.6274	0.0008

5.2. Experiment 2: Intra-rater segmentation consistency

The second experiment attempts to quantify intra-rater consistency, comparing the two cycles of segmentations of each rater, separately for IML and manual (Table 3, Fig.11). No

significant difference was found between manual and IML-assisted segmentations for any of the cohorts, except spleen where segmentations using the method had higher mean overlap.

Additional analysis of DSC relative to ground truth (Table 2) found a significant difference in only one of the raters when using the IML method, while revealing a significant difference in 4/8 raters for manual annotations. Furthermore, there was no significant difference when using the IML method for any of the two raters for individual sub-regions of glioblastoma. The same tests for manual annotations revealed a significant difference in all sub-regions, except ET in one of the raters.

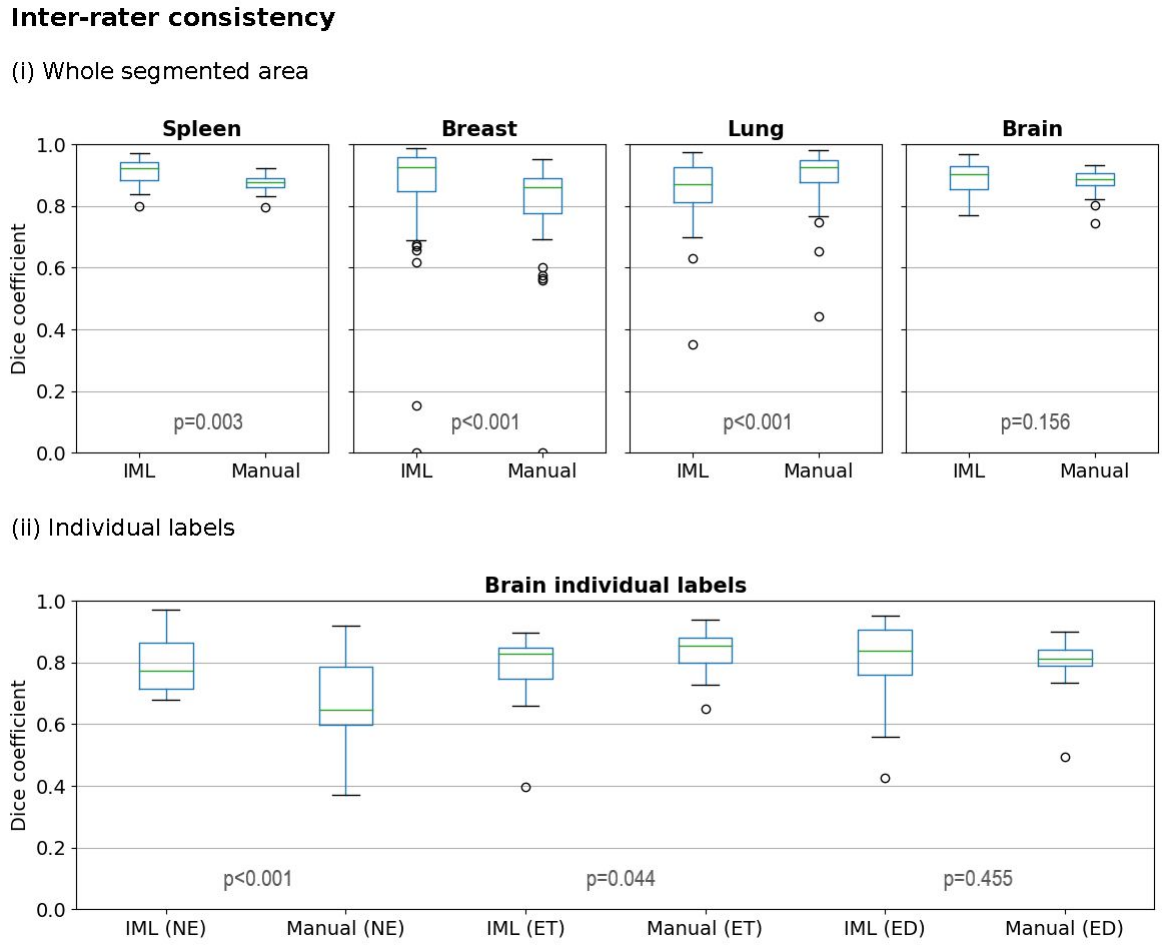

Figure 12: Dice coefficient, for inter-rater consistency, between segmentations of different raters, for: (i) all individual labels representing different areas of the structure counted as one, (ii) the individual areas of glioblastomas.

5.3. Experiment 3: Inter-rater segmentation consistency

In the last experiment, inter-rater consistency of IML and manual segmentations was calculated and compared (Table 3, Fig.12). There was a significant difference for spleen, breast, lung, and the NE and ET glioblastoma sub-regions. From those, the method achieved a higher overlap for spleen, breast, and the NE glioblastoma sub-region. Conversely, manual segmentations had higher agreement for lung and the ET region.

Table 3: Results for all three experiments. Experiment 1: Performance is calculated as dice coefficient relative to ground truth. p-values are a result of paired comparisons between the highest scoring IML-assisted and manual segmentations for each rater and each case. Correlation coefficient is calculated between the resultant and ground truth volumes. Experiment 2: Values indicate overlap between the first and second cycle of each rater. $p<0.05$ indicates a significant difference between the results of IML and manual segmentations. Experiment 3: Values indicate overlap between segmentations of different raters. $p<0.05$ indicates a significant difference in the inter-rater results for the respective cohort.

	Experiment 1							**Experiment 2**			**Experiment 3**		
	Mean Overlap			Time		Corr. Coef.		Intra-rater mean			Inter-rater mean		
Label	IML	Manual	p	IML	Manual	IML	Manual	IML	Manual	p	IML	Manual	p
Spleen													
-	0.91	0.87	$<10^{-3}$	66s	100s	0.99	0.99	0.91	0.89	0.003	0.91	0.87	0.003
Breast													
-	0.84	0.82	$<10^{-3}$	19s	70s	0.98	0.95	0.88	0.9	0.37	0.86	0.81	$<10^{-3}$
Lung													
-	0.78	0.83	$<10^{-3}$	93s	125s	0.96	0.97	0.96	0.95	0.116	0.85	0.89	$<10^{-3}$
Brain													
WT	0.94	0.92	0.259	21m	60m	1	0.98	0.91	0.89	0.61	0.89	0.88	0.156
NE	0.81	0.79	0.183			0.97	0.95	0.85	0.86	0.32	0.79	0.67	$<10^{-3}$
ET	0.85	0.87	0.09			0.96	0.98	0.85	0.88	0.116	0.79	0.84	0.044
ED	0.88	0.87	0.657			1	0.98	0.85	0.83	0.788	0.81	0.8	0.455

6. Discussion

In this study, a general-purpose, easy-to-use, and fast IML-based segmentation method was presented that can be applied in a multitude of research applications without requiring any adaptations to different domains or training of users. The method takes as input co-registered images and quick user drawings, to create AGD maps and train an ensemble of SVMs, used for segmenting the whole scan. The performance of the method was evaluated on solid structures across different cohorts, image modalities, and anatomical sites.

The method utilizes the power of ML; however, it mitigates one of its known weaknesses, i.e., the need for extensive training and lack of reproducibility on new datasets. By virtue of being trained interactively, segmentation models are optimal for the specific individual's scans. Additional benefits include the ability of the method to be parallelized and low hardware requirements. The disadvantage of this approach is that it is not fully automated.

Quantitative evaluation showed great promise for the applicability of the method in various structures relevant to medical research. Accuracy and inter-rater agreement were comparable to manual segmentation, while intra-rater agreement was high, indicating that the method is stable. Volumetric errors were mostly systematic, indicating that results can be improved through further iterations or volumetric operations like shrinking/expanding. The method was also shown to be fast and not require excessive interaction.

The presented method can be utilized in a vast array of applications, because fully automatic segmentation has not received widespread clinical adoption yet. Training of automatic methods requires large datasets that are difficult to create and obtain. Additionally, deep learning methods are often inaccessible to clinicians, because they require specialized hardware and don't have a standardized distribution method yet. Furthermore, automatic

methods target specific problems, which means that it would take years, expensive acquisition trials, and vast inter-institutional cooperation to build robust models for every structure of the body. Some automatic approaches also face the problem of domain adaptation, meaning that the trained models have trouble segmenting images acquired from machines with different intensity and noise profiles than the ones they were trained on. As a result, interactive segmentation is currently highly relevant and will probably continue to be, to some extent, in the future. Lastly, medical image segmentation is complicated, because label assignment often relies on advanced medical knowledge and patient information, especially for cancerous structures. Consequently, automatic methods will probably never become synonymous with ground truth and there will be a need for quick expert corrections, potentially with interactive methods that use a segmentation produced automatically alongside input from a clinician to improve the result.

ITK-SNAP [14] also provides a method for interactive segmentation. However, it requires users to follow a more complex protocol to achieve multi-label segmentation. The user first provides quick drawings for the different ROIs and trains a model. Afterwards and separately for each class, the user must place seeds and evolve a contour. According to their evaluation on high-grade glioblastomas, also on the BraTS dataset, there is a lower mean agreement with the ground truth for ITK-SNAP in the regions this thesis also evaluated, particularly ET ($Dice_{IML}/Dice_{ITK\text{-}SNAP}$=0.85/0.69) and WT ($Dice_{IML}/Dice_{ITK\text{-}SNAP}$=0.94/0.85). Average user interaction time was also lower for the method presented here ($Time_{IML}/Time_{ITK\text{-}SNAP}$=21min/27.8min). 3D Slicer's "grow from seeds" effect follows a workflow similar to the one that was presented here, but it can only support one image as input.

Future research can improve this method on multiple fronts. Advanced ML techniques, such as semi-supervised learning, can potentially increase the accuracy and consistency of the results. Transfer learning could expand the range of tasks to non-solid structures, such as

brain lesions. If a specific task is targeted, pre-trained population-derived models, atlases, and specialized preprocessing techniques can potentially aid in producing better segmentations. Furthermore, a prospective dataset, especially one acquired under different acquisition settings, would lend further validity to the method.

The results showed that the method has accuracy and inter-rater consistency on par with manual segmentation across different solid anatomical structures and modalities. Additionally, the method showed high intra-rater consistency and minimized user interaction.

7. Conclusion

The purpose of this thesis was to compare the proposed method with manual segmentation in an experimental setting that included segmentation of spleens and brain, breast, and lung tumors. Various tests concluded that the method is comparable to manual segmentation in accuracy, intra- and inter-rater consistency in most situations, while even outperforming manual segmentation in some others. Additionally, the method was always faster, making it suitable for applications that require segmenting many cases.

8. References

[1] D. L. Pham, C. Xu, and J. L. Prince, “Current methods in medical image segmentation,” *Annu. Rev. Biomed. Eng.*, vol. 2, pp. 315–337, 2000, doi: 10.1146/annurev.bioeng.2.1.315.

[2] N. Sharma and L. M. Aggarwal, “Automated medical image segmentation techniques,” *J. Med. Phys.*, vol. 35, no. 1, pp. 3–14, Jan. 2010, doi: 10.4103/0971-6203.58777.

[3] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/tmi.2014.2377694.

[4] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci Data*, vol. 4, p. 170117, 2017, doi: 10.1038/sdata.2017.117.

[5] S. Golemati and K. Nikita, Eds., *Cardiovascular Computing-Methodologies and Clinical Applications*. Springer, 2020.

[6] S. Rathore *et al.*, “Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning,” *J Med Imaging (Bellingham)*, vol. 5, no. 2, p. 021219, Apr. 2018, doi: 10.1117/1.Jmi.5.2.021219.

[7] B. Sahiner *et al.*, *The effect of nodule segmentation on the accuracy of computerized lung nodule detection on CT scans: comparison on a data set annotated by multiple radiologists*, vol. 6514. SPIE, 2007, p. MI.

[8] S. Golemati, J. Stoitsis, E. G. Sifakis, T. Balkizas, and K. S. Nikita, “Using the Hough transform to segment ultrasound images of longitudinal and transverse sections of the carotid artery,” *Ultrasound Med. Biol.*, vol. 33, no. 12, pp. 1918–1932, Dec. 2007, doi: 10.1016/j.ultrasmedbio.2007.05.021.

[9] A. S. Panayides *et al.*, “AI in Medical Imaging Informatics: Current Challenges and Future Directions,” *IEEE Journal of Biomedical and Health Informatics*. pp. 1–1, 2020, doi: 10.1109/jbhi.2020.2991043.

[10] J. H. Thrall *et al.*, “Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success,” *J. Am. Coll. Radiol.*, vol. 15, no. 3, Part B, pp. 504–508, Mar. 2018, doi: 10.1016/j.jacr.2017.12.026.

[11] P. A. Yushkevich *et al.*, “User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability,” *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006, doi: 10.1016/j.neuroimage.2006.01.015.

[12] S. Bakas *et al.*, “GLISTRboost: Combining Multimodal MRI Segmentation, Registration, and Biophysical Tumor Growth Modeling with Gradient Boosting Machines for Glioma Segmentation,” *Brainlesion*, vol. 9556, pp. 144–155, 2016, doi: 10.1007/978-3-319-30858-6_1.

[13] K. Zeng *et al.*, “Segmentation of Gliomas in Pre-operative and Post-operative Multimodal Magnetic Resonance Imaging Volumes Based on a Hybrid Generative-Discriminative Framework,” *Brainlesion*, vol. 10154, pp. 184–194, Oct. 2016, doi: 10.1007/978-3-319-55524-9_18.

[14] P. A. Yushkevich *et al.*, “User-Guided Segmentation of Multi-modality Medical Imaging Datasets with ITK-SNAP,” *Neuroinformatics*, vol. 17, no. 1, pp. 83–102, Jan. 2019, doi: 10.1007/s12021-018-9385-x.

[15] A. Fedorov *et al.*, “3D Slicer as an image computing platform for the Quantitative Imaging Network,” no. 1873–5894 (Electronic), 2012.

[16] B. Gaonkar *et al.*, “Automated Tumor Volumetry Using Computer-Aided Image Segmentation,” *Acad. Radiol.*, vol. 22, no. 5, pp. 653–661, 2015, doi: 10.1016/j.acra.2015.01.005.

[17] “Diagnostic imaging,” *World Health Organization*. https://www.who.int/diagnostic_imaging/en/ (accessed Jul. 21, 2020).
[18] E. Friedberg *et al.*, “Unifying the Silos of Subspecialized Radiology: The Essential Role of the General Radiologist,” *J. Am. Coll. Radiol.*, vol. 15, no. 8, pp. 1158–1163, Aug. 2018, doi: 10.1016/j.jacr.2018.05.016.
[19] “Imaging and radiology,” *medlineplus.gov*. https://medlineplus.gov/ency/article/007451.htm (accessed Jul. 20, 2020).
[20] T. P. Murphy, “Introduction to clinical interventional radiology,” *Semin. Intervent. Radiol.*, vol. 22, no. 1, pp. 3–5, Mar. 2005, doi: 10.1055/s-2005-869569.
[21] T. Kurki, N. Lundbom, and S. Valtonen, “Tissue characterisation of intracranial tumours: the value of magnetisation transfer and conventional MRI,” *Neuroradiology*, vol. 37, no. 7, pp. 515–521, Oct. 1995, doi: 10.1007/BF00593707.
[22] P. Mildenberger, M. Eichelberg, and E. Martin, “Introduction to the DICOM standard,” *Eur. Radiol.*, vol. 12, no. 4, pp. 920–927, Apr. 2002, doi: 10.1007/s003300101100.
[23] I. Wolf *et al.*, “The medical imaging interaction toolkit,” *Med. Image Anal.*, vol. 9, no. 6, pp. 594–604, Dec. 2005, doi: 10.1016/j.media.2005.04.005.
[24] GCO, “All Cancers - GCO,” *International Agency for Research on Cancer*. https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf (accessed Jul. 27, 2020).
[25] P. Anand *et al.*, “Cancer is a preventable disease that requires major lifestyle changes,” *Pharm. Res.*, vol. 25, no. 9, pp. 2097–2116, Sep. 2008, doi: 10.1007/s11095-008-9661-9.
[26] “Types of cancer,” *Cancer Research UK*. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer (accessed Jul. 27, 2020).
[27] “What is Cancer?,” *cancer.net*. https://www.cancer.net/navigating-cancer-care/cancer-basics/what-cancer (accessed Sep. 08, 2020).
[28] A. Gupta and T. Dwivedi, “A Simplified Overview of World Health Organization Classification Update of Central Nervous System Tumors 2016,” *J. Neurosci. Rural Pract.*, vol. 8, no. 4, pp. 629–641, Oct. 2017, doi: 10.4103/jnrp.jnrp_168_17.
[29] D. N. Louis *et al.*, “WHO classification and grading of tumours of the central nervous system,” *WHO Classification of Tumours of the Central Nervous System. 4th ed. , Revised. Lyon: International Agency for Research Centre*, pp. 12–13, 2016.
[30] A. Brodbelt, D. Greenberg, T. Winters, M. Williams, S. Vernon, and V. P. Collins, “Glioblastoma in England: 2007–2011,” *Eur. J. Cancer*, vol. 51, no. 4, pp. 533–542, Mar. 2015, doi: 10.1016/j.ejca.2014.12.014.
[31] L. Sun, S. Zhang, H. Chen, and L. Luo, “Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning,” *Front. Neurosci.*, vol. 13, p. 810, Aug. 2019, doi: 10.3389/fnins.2019.00810.
[32] S. Bakas *et al.*, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” *arXiv*, p. arXiv:1811.02629, 2018, [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv181102629B.
[33] O. Eidel *et al.*, “Tumor Infiltration in Enhancing and Non-Enhancing Parts of Glioblastoma: A Correlation with Histopathology,” *PLoS One*, vol. 12, no. 1, p. e0169292, Jan. 2017, doi: 10.1371/journal.pone.0169292.
[34] S. G. Armato 3rd, M. L. Giger, and H. MacMahon, “Automated detection of lung nodules in CT scans: preliminary results,” *Med. Phys.*, vol. 28, no. 8, pp. 1552–1561, Aug. 2001, doi: 10.1118/1.1387272.
[35] J. Wei, Y. Hagihara, A. Shimizu, and H. Kobatake, “Optimal image feature set for

detecting lung nodules on chest X-ray images," in *CARS 2002 Computer Assisted Radiology and Surgery*, 2002, pp. 706–711, doi: 10.1007/978-3-642-56168-9_118.
[36] M. Woźniak, D. Połap, G. Capizzi, G. L. Sciuto, L. Kośmider, and K. Frankiewicz, "Small lung nodules detection based on local variance analysis and probabilistic neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 173–180, Jul. 2018, doi: 10.1016/j.cmpb.2018.04.025.
[37] C. A. Owens *et al.*, "Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer," *PLoS One*, vol. 13, no. 10, p. e0205003, Oct. 2018, doi: 10.1371/journal.pone.0205003.
[38] C. E. DeSantis *et al.*, "Breast cancer statistics, 2019," *CA Cancer J. Clin.*, vol. 69, no. 6, pp. 438–451, Nov. 2019, doi: 10.3322/caac.21583.
[39] S. Park *et al.*, "Changes in Noninvasive Liver Fibrosis Indices and Spleen Size During Chemotherapy: Potential Markers for Oxaliplatin-Induced Sinusoidal Obstruction Syndrome," *Medicine* , vol. 95, no. 2, p. e2454, Jan. 2016, doi: 10.1097/MD.0000000000002454.
[40] F. A. Fasola and A. J. Adekanmi, "HAEMATOLOGICAL PROFILE AND BLOOD TRANSFUSION PATTERN OF PATIENTS WITH SICKLE CELL ANAEMIA VARY WITH SPLEEN SIZE," *Ann Ib Postgrad Med*, vol. 17, no. 1, pp. 30–38, Jun. 2019, [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31768154.
[41] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.
[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27–21, 2011.
[43] R. Filipovych and C. Davatzikos, "Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI)," *Neuroimage*, vol. 55, no. 3, pp. 1109–1119, Apr. 2011, doi: 10.1016/j.neuroimage.2010.12.066.
[44] M. Ghafoorian *et al.*, "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation," in *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017*, 2017, pp. 516–524, doi: 10.1007/978-3-319-66179-7_59.
[45] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, "Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols," *IEEE Trans. Med. Imaging*, vol. 34, no. 5, pp. 1018–1030, May 2015, doi: 10.1109/TMI.2014.2366792.
[46] H.-C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.
[47] N. Heller *et al.*, "The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes," *arXiv [q-bio.QM]*, Mar. 31, 2019.
[48] Y. Sun *et al.*, "Multi-Site Infant Brain Segmentation Algorithms: The iSeg-2019 Challenge," *arXiv [eess.IV]*, Jul. 04, 2020.
[49] G. Litjens *et al.*, "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge," *Med. Image Anal.*, vol. 18, no. 2, pp. 359–373, Feb. 2014, doi: 10.1016/j.media.2013.12.002.
[50] B. Boyer, C. Balleyguier, O. Granat, and C. Pharaboz, "CAD in questions/answers Review of the literature," *Eur. J. Radiol.*, vol. 69, no. 1, pp. 24–33, Jan. 2009, doi: 10.1016/j.ejrad.2008.07.042.
[51] J. Firth-Cozens, "Doctors, their wellbeing, and their stress," *BMJ*, vol. 326, no. 7391, pp. 670–671, Mar. 2003, doi: 10.1136/bmj.326.7391.670.
[52] S. G. Mougiakakou, S. Golemati, I. Gousias, A. N. Nicolaides, and K. S. Nikita, "Computer-aided diagnosis of carotid atherosclerosis based on ultrasound image

statistics, Laws' texture and neural networks," *Ultrasound in Medicine & Biology*, vol. 33, no. 1. pp. 26–36, 2007, doi: 10.1016/j.ultrasmedbio.2006.07.032.
[53] M. Gletsos, S. G. Mougiakakou, G. K. Matsopoulos, K. S. Nikita, A. S. Nikita, and D. Kelekis, "A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 153–162, Sep. 2003, doi: 10.1109/titb.2003.813793.
[54] J. Stoitsis, I. Valavanis, S. G. Mougiakakou, S. Golemati, A. Nikita, and K. S. Nikita, "Computer aided diagnosis based on medical image processing and artificial intelligence methods," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 569, no. 2. pp. 591–595, 2006, doi: 10.1016/j.nima.2006.08.134.
[55] K. S. Nikita, "Atherosclerosis: the evolving role of vascular image analysis," *Comput. Med. Imaging Graph.*, vol. 37, no. 1, pp. 1–3, Jan. 2013, doi: 10.1016/j.compmedimag.2012.12.001.
[56] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Syst. Appl.*, vol. 138, p. 112821, Dec. 2019, doi: 10.1016/j.eswa.2019.112821.
[57] K. Nikita, D. Koutsouris, and S. Pavlopoulos, *Ιατρικά απεικονιστικά συστήματα*. ΤΖΙΟΛΑΣ, 2004.
[58] N. N. Tsiaparas, S. Golemati, I. Andreadis, J. S. Stoitsis, I. Valavanis, and K. S. Nikita, "Comparison of multiresolution features for texture classification of carotid atherosclerosis from B-mode ultrasound," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 130–137, Jan. 2011, doi: 10.1109/TITB.2010.2091511.
[59] S. G. Mougiakakou, I. Valavanis, K. S. Nikita, A. Nikita, and D. Kelekis, "Characterization of CT liver lesions based on texture features and a multiple neural network classification scheme," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, 2003, vol. 2, pp. 1287–1290 Vol.2, doi: 10.1109/IEMBS.2003.1279504.
[60] P. Asvestas, G. K. Matsopoulos, and K. S. Nikita, "A Power Differentiation Method of Fractal Dimension Estimation for 2-D Signals," *J. Vis. Commun. Image Represent.*, vol. 9, no. 4, pp. 392–400, Dec. 1998, doi: 10.1006/jvci.1998.0394.
[61] J. J. M. van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.
[62] S. E. Chatzistergos, I. Andreadis, and K. S. Nikita, "Identification of architectural distortions in mammograms using local binary patterns and radial lengths through an exhaustive evaluation framework," *Expert Syst.*, vol. 35, no. 4, p. e12281, Aug. 2018, doi: 10.1111/exsy.12281.
[63] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013, [Online]. Available: https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s10115-012-0487-8&casa_token=8S2e0NubWboAAAAA:Flb3uuxUkyfq-1gI8RrzXJs9FnHuMs1Q2lumEQHQXJ6-BQBiEhCpK8tWm2fPBHQd8CrSLh4KwY9tDofH.
[64] B. Gaonkar, L. Shu, G. Hermosillo, and Y. Zhan, "Adaptive geodesic transform for segmentation of vertebrae on CT images," in *Medical Imaging 2014: Computer-Aided Diagnosis*, Mar. 2014, vol. 9035, p. 903516, doi: 10.1117/12.2043527.
[65] K. H. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad. Radiol.*, vol. 11, no. 2, pp. 178–189, Feb. 2004, doi: 10.1016/s1076-6332(03)00671-8.

[66] S. M. Stigler, “Francis Galton’s Account of the Invention of Correlation,” *Stat. Sci.*, vol. 4, no. 2, pp. 73–79, 1989.
[67] C. Davatzikos *et al.*, “Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome,” *J Med Imaging (Bellingham)*, vol. 5, no. 1, p. 011018, Jan. 2018, doi: 10.1117/1.Jmi.5.1.011018.
[68] M. McCormick, X. Liu, J. Jomier, C. Marion, and L. Ibanez, “ITK: enabling reproducible research and open science,” *Front. Neuroinform.*, vol. 8, p. 13, Feb. 2014, doi: 10.3389/fninf.2014.00013.
[69] T. S. Yoo *et al.*, “Engineering and algorithm design for an image processing Api: a technical report on ITK--the Insight Toolkit,” *Stud. Health Technol. Inform.*, vol. 85, pp. 586–592, 2002, [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15458157.
[70] BRADSKI and G, “The OpenCV library,” *Dr Dobb’s J. Software Tools*, vol. 25, pp. 120–125, 2000, Accessed: Jun. 16, 2020. [Online]. Available: https://ci.nii.ac.jp/naid/10028167478/.
[71] W. McKinney and Others, “Data structures for statistical computing in python,” in *Proceedings of the 9th Python in Science Conference*, 2010, vol. 445, pp. 51–56, [Online]. Available: http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.
[72] T. E. Oliphant, *A guide to NumPy*, vol. 1. Trelgol Publishing USA, 2006.
[73] P. Virtanen *et al.*, “Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 352–352, Mar. 2020, doi: 10.1038/s41592-020-0772-5.
[74] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.
[75] A. L. Simpson *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” *arXiv*, p. arXiv:1902.09063, 2019, [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190209063S.
[76] H. J. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nat. Commun.*, vol. 5, p. 4006, 2014, doi: 10.1038/ncomms5006.
[77] H. J. W. L. Aerts *et al.*, “Data From NSCLC-Radiomics [Data set],” *The Cancer Imaging Archive.*, 2019.
[78] K. Clark *et al.*, “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository,” *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013, doi: 10.1007/s10278-013-9622-7.
[79] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.