University of Technology of Compiègne

Doctoral Thesis

Medical Image Segmentation with Belief Function Theory and Deep Learning

Author:
Ling Huang

Supervisors:
Prof. Thierry Denoeux
Prof. Su Ruan

Jury:

Prof. Franck DAVOINE	UTC	President of jury
Prof. Sylvie LE HÉGARAT	University of Paris-Saclay	Reviewer
Prof. David MERCIER	University of Artois	Reviewer
Asso Prof. Emmanuel RAMASSO	ENSMM School	Examiner
Prof. Su RUAN	University of Rouen	Co-supervisor
Prof. Thierry DENOEUX	UTC	Supervisor

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

CID Team (Knowledge, Uncertainty, Data)
Heudiasyc Laboratory

March 10, 2023

UNIVERSITY OF TECHNOLOGY OF COMPIÈGNE

Abstract

CID Team (Knowledge, Uncertainty, Data)
Heudiasyc Laboratory

Doctor of Philosophy

Medical Image Segmentation with Belief Function Theory and Deep Learning

by Ling HUANG

Deep learning has shown promising contributions in medical image segmentation with powerful learning and feature representation abilities. However, it has limitations for reasoning with and combining imperfect (imprecise, uncertain, and partial) information. In this thesis, we study medical image segmentation approaches with belief function theory and deep learning, specifically focusing on information modeling and fusion based on uncertain evidence.

First, we review existing belief function theory-based medical image segmentation methods and discuss their advantages and challenges. Second, we present a semi-supervised medical image segmentation framework to decrease the uncertainty caused by the lack of annotations with evidential segmentation and evidence fusion. Third, we compare two evidential classifiers, evidential neural network and radial basis function network, and show the effectiveness of belief function theory in uncertainty quantification; we use the two evidential classifiers with deep neural networks to construct deep evidential models for lymphoma segmentation. Fourth, we present a multimodal medical image fusion framework taking into account the reliability of each MR image source when performing different segmentation tasks using mass functions and contextual discounting.

Keywords: Belief function theory, Deep learning, Medical image analysis, Semi-supervised learning, Uncertainty quantification, Information fusion

Acknowledgements

First and foremost, I am extremely grateful to my two supervisors, Prof. Thierry Denœux from UTC and Prof. Su Ruan from University of Rouen Nornamdie for their invaluable advice and continuous support during my Ph.D. study. It is my great luck to be supervised by Prof. Thierry Denœux and Prof.Su Ruan. Without their assistance and dedicated involvement, this thesis would have never been accomplished.

I would like to express my great gratitude to my supervisor, Prof. Thierry Denœux, for his constructive guidance during my PhD study. As a scrupulous and creative scholar working in the field of uncertainty modeling, belief function theory, information fusion, Prof. Thierry Denœux has taught me tremendous expertise. He is the leader of my scientific research career and is always encourage me to do a better work by his insightful comments. It is him that helped me to build my inner confidence about scientific research and helped me to find a solid, long-term and sustainable way to move forward. He has set a very good example to me about being a conscientious and professional researcher. I am also very grateful for his understanding and supports for every important moment during my Ph.D.

I would like to thank my co-supervisor, Prof. Su Ruan, for her nuanced and constructive guidance. Thanks to her strong expertise and experience in medical image analysis and machine learning, my research background has been enhanced to face future challenges. I am impressed by her passion for work and the quality of hardworking. I appreciate her cheerful encouragement when I face difficulties in research. I also appreciate her help and support in my life during my stay in Rouen as well as in Compiègne.

I would also like to thanks all the jury members, Prof. Sylvie Le Hegarat-MASCL from Université Paris-Sud, and Prof. David MERCIER from Université of Artois, Prof. Emmanuel Ramasso from ENSMM School for spending their valuable time to review my thesis manuscript and give me valuable comments. I would like to give a special thanks to Dr. Franck DAVOINE from UTC for his help during my Ph.D. and also for accepting to be the president of jury in my thesis defense.

I would to thank my colleagues from Heudiasyc in Compiègne with whom I spent a huge amount of time during office hours and beyond. Thanks to all of you for the moments we shared. I would like to say thanks to my colleagues from LITIS Quantif in Rouen for their help during my stay in Rouen. I would also like to thank my friends from China and France, for their warm help when I need.

I would like to gratefully acknowledge the support provided for this thesis under the framework of the Labex MS2T. I would also like to thank the Chinese Scholarship Council for supporting my study in France.

I would like to express great thanks to my family, in particular to my beloved parents, my brother and my sister-in-law, for their continuous support and encouragement throughout the years of my study.

Last but not least, I would also like to thank myself for dedicating my gold age to finishing this thesis. All efforts are worth it.

List of Publications

Journal papers

- **Ling Huang**, Su Ruan and Thierry Denœux. ”Application of belief functions to medical image segmentation: A review.” *Information Fusion*, 91 (2023), pp. 737-756, https://doi.org/10.1016/j.inffus.2022.11.008.
- **Ling Huang**, Su Ruan, Pierre Decazes, Thierry Denœux. ”Lymphoma segmentation from 3D PET-CT images using a deep evidential network.” *International Journal of Approximate Reasoning*, 149 (2022), pp. 39-60, https://doi.org/10.1016/j.ijar.2022.06.007.
- **Ling. Huang**, Su Ruan and Thierry Denœux. ”Semi-Supervised Multiple Evidence Fusion for Brain Tumor Segmentation.” Neurocomputing. (Accepted).

Conferences papers

- **Ling Huang**, Thierry Denœux, Pierre Vera, Su Ruan. ”Evidence fusion with contextual discounting for multi-modality medical image segmentation.” *International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI). Springer, Cham, 2022.
- **Ling Huang**, Su Ruan, Pierre Decazes, Thierry Denœux. ”Evidential segmentation of 3D PET/CT images.” *International Conference on Belief Functions* (BELIEF). Springer, Cham, 2021.
- **Ling Huang**, Su Ruan, and Thierry Denœux. ”Belief function-based semi-supervised learning for brain tumor segmentation.” *18th International Symposium on Biomedical Imaging* (ISBI). IEEE, 2021.
- **Ling Huang**, Su Ruan, and Thierry Denœux. ”Covid-19 classification with deep neural network and belief functions.” *The Fifth International Conference on Biological Information and Biomedical Engineering* (BIBE). 2021.

Workshop paper

- **Ling Huang**, Thierry Denœux, David Tonnelet, Pierre Decazes and Su Ruan. ”Deep PET/CT fusion with Dempster-Shafer theory for lymphoma segmentation.” *18th International Workshop on Machine Learning in Medical Imaging* (MLMI). Springer, Cham, 2021.

Available codes

- The python code using open-source software library Pytorch and MONAI can be downloaded at https://github.com/iWeisskohl.

Contents

List of Symbols

Ω	frame of discernment
m	mass function
Bel	belief function
Pl	plausibility function
pl	contour function
$BetP$	Pignistic probability
$\oplus$	orthogonal sum
κ	degree of conflict
β	degree of belief that the source mass function is reliable
${}^{\beta}m$	discounted mass function
$\boldsymbol{\beta}$	a vector of reliability values
$m_?$	vacuous mass function
p_i	a prototype in an ENN or RBF classifier
d_i	distance between x and prototype p_i
s_i	distance-based support between x and prototype p_i
$\underline{E}_m(u), \overline{E}_m(u)$	lower, upper expected utilities associated to m
$pCNN, mENN$	segmentation outputs of probability module and mass function module
S, G	output results and ground truth
x_t	the transformed data based on input x
ν_i	weight of the connection between hidden unit i and output unit
w_i	as weights of evidence for class 1 or 2, depending on the sign of ν_i
λ	regularization coefficient
R	regularizer

Introduction

research background and challenges

In clinical routine, the segmentation or delineation of the target region is performed manually by physicians. However, the manual segmentation operation has limited efficiency and accuracy. First, image segmentation is time-consuming, especially with 3D imagesv. It leads to a waste of medical resources. Second, the segmentation performance is limited by the quality of medical images, the difficulty of the disease, and the domain knowledge of the experts. Thus, the study on automatic medical image segmentation methods with powerful artificial intelligence technologies is necessary to address the problem of limited medical resources and improve the segmentation performance.

Classical medical image segmentation approaches [Batenburg et al., 2009; Kimmel, 2003; Salvador et al., 2004; Strobl et al., 2007; Kleinbaum et al., 1996] focus on low-level feature analysis, e.g., gray and textual features or hand-crafted features. These approaches have limitations in terms of segmentation accuracy, efficiency, and reliability, which creates a big gap between experimental performance and clinical application. More recently, the success of deep learning in the medical domain brought a lot of contributions to medical image segmentation tasks [Ronneberger et al., 2015; Myronenko, 2018; Isensee et al., 2018; Bahdanau et al., 2014]. Deep learning solves the segmentation efficiency problem, allowing for large-scale medical image segmentation, and contributes greatly to its accuracy. However, large-scale labeled training data are needed to reach a satisfying segmentation performance. Obtaining precisely annotated data is particularly challenging in the medical image segmentation domain, which has become one of the bottlenecks of learning-based segmentation approaches.

Despite the excellent performance of deep learning-based medical image segmentation methods, doubts about the reliability of the segmentation results still remain [Hüllermeier et al., 2021], which explains why their application to therapeutic decision-making for complex oncological cases is still limited. A reliable segmentation model should be well calibrated, i.e., its confidence should match its accuracy. Therefore, a trustworthy representation of uncertainty is desirable and should be considered a key feature of any deep learning method, especially in safety-critical application domains, e.g., medical image segmentation. In general, deep models have two sources of uncertainty: aleatory uncertainty and epistemic uncertainty [Hora, 1996; Der Kiureghian et al., 2009]. Aleatory uncertainty refers to the notion of randomness, i.e., the variability in an experimental outcome due to inherently random effects. In contrast, epistemic uncertainty refers to uncertainty caused by a lack of knowledge (ignorance) about the best model, i.e., the ignorance of the learning algorithm or decision-maker. As opposed to uncertainty caused by randomness, uncertainty caused by ignorance can be reduced based on additional information or the design of a suitable learning algorithm. In medical image segmentation, the uncertainty can be decomposed into three levels [Lakshminarayanan et al., 2017]. *Pixel/voxel-level* uncertainty is useful for interaction with physicians by providing additional guidance for correcting segmentation results. *Instance-level* is the uncertainty aggregate by pixel/voxel-level uncertainty, which can be used to reduce the false discovery rate. *Subject-level* uncertainty offers information on whether the segmentation model is reliable. Early approaches to quantify the segmentation uncertainty were based on Bayesian theory [Hinton et al., 1993; MacKay, 1992]. The popularity of deep segmentation models has revived research on model uncertainty estimation and has given rise to specific methods such as variational dropout [Gal et al., 2016; Tran et al., 2019], and model ensembles [Lakshminarayanan et al., 2017; Rupprecht et al., 2017]. However, probabilistic segmentation models capture knowledge in terms of a single probability distribution and cannot distinguish between aleatory and epistemic uncertainty, which limits the exploitation of the results.

Furthermore, single-modality biomedical data does not provide a complete representation of disease information, and information sources are usually imperfect. The advances in medical imaging machines and technologies now allow us to obtain medical images with several modalities, such as Magnetic Resonance Imaging (MRI)/Positron Emission Tomography (PET), multi-sequence MRI, or PET/Computed Tomography (CT). The different modalities can provide complementary information about cancer and other abnormalities in the human body. Thus, the fusion of multiple information is vital to improve the accuracy of diagnosis and help radiotherapy. The success of information fusion depends on how well the fused knowledge represents reality [White, 1991]. More specifically, it depends on how adequate the input information is, how accurate and appropriate prior knowledge is, and on the quality of the uncertainty model used [Rogova et al., 2004]. Though deep learning has shown promising performance in information representation, modeling imperfect data and combining unreliable sources need further study.

In this thesis, we address medical image segmentation using belief function theory (BFT) [Dempster, 1967; Shafer, 1976; Denœux et al., 2020] and deep neural networks, with a focus on uncertainty quantification and information fusion. BFT is a theoretical framework for modeling, reasoning with and combining imperfect (imprecise, uncertain, and partial) information. With BFT, we can quantify epistemic uncertainty directly and further explore the possibility of improving the model reliability based on the quantified uncertainty. Besides uncertainty quantification, Dempster's rule of BFT offers a way to fuse multiple uncertain information, e.g., multimodal medical images, to achieve more promising segmentation performance [Lian et al., 2018a; Ghesu et al., 2021]. Motivated by the high expressivity of BFT, the goals of this thesis are 1) to develop evidential medical image segmentation models in the framework of BFT and deep neural networks, with the cooperation of powerful learning algorithms such as semi-supervised learning, 2) to study the performance of deep evidential segmentation with uncertainty quantification and 3) to study the effectiveness of BFT for constructing deep multimodal segmentation model.

CONTRIBUTIONS OF THE THESIS

The contributions of this thesis are organized in three parts:

- **Semi-supervised medical image segmentation**: The main idea of this part is to address the annotation limitation with the design of semi-supervised learning and decrease the uncertainty caused by the lack of annotations with evidential segmentation and evidence fusion. We first propose a semi-supervised learning algorithm based on an image transformation strategy by producing pseudo labels for unannotated data. Then a probabilistic-based deep neural network and a BFT-based evidential neural network are used to compute segmentation results with uncertainty quantified by probabilities and mass functions. Finally, these probabilities and mass functions are combined by Dempster's rule.

- **Uncertainty quantification in medical image segmentation**: A reasonable and reliable quantification of segmentation uncertainty is important to optimize the segmentation framework and further improve performance. In this part, an automatic evidential segmentation model based on BFT and deep learning is proposed to segment lymphomas from 3D PET-CT images, which not only focuses on lymphoma segmentation accuracy but also on uncertainty quantification using belief functions. The model is composed of a deep feature-extraction module and an evidential layer. The feature extraction module uses an encoder-decoder framework to extract semantic feature vectors from 3D inputs. The evidential layer then uses prototypes in the feature space to compute a belief function at each voxel, quantifying the uncertainty about the presence or absence of lymphoma at this location. Two evidential layers are compared based on different ways of using distances to prototypes for computing mass functions.

- **Multimodal medical image segmentation with contextual discounting**: Different modality biomedical data have different reliabilities to segment specific diseases. Thus, quantifying context-based modality-level uncertainty and combining multimodal information is essential to reach a reliable decision and explain the decision. In this part, we construct a multi-MR image brain tumor segmentation framework by modeling the reliability of each source of MR images when doing different segmentation tasks using mass functions and contextual discounting. Different modality MR images are taken as an independent source of input. One feature extraction and one evidential segmentation module are used for each modality input to assign each voxel a mass function. Then a contextual discounting layer is designed to take into account the reliability of the single modality MR images when classifying different diseases. Finally, the discounted evidence from different modalities is combined by Dempster's rule to obtain a final segmentation.

LAYOUT OF THE THESIS

This thesis is structured in two parts and seven chapters:

- Part I introduces the theoretical background and general context that supports the thesis. Chapter 1 first introduces medical image segmentation and its significance in clinical treatment and then introduces classical deep segmentation methods and their limitations. Chapter 2 recalls the fundamentals of belief function theory. It first describes how imperfect (uncertain and imprecise) information can be modeled and combined in the framework of BFT, and then introduces the basic belief assignment methods that generate mass function for medical image segmentation. Chapter 3 comprehensively reviews existing medical image segmentation methods using BFT.

- Part II is devoted to our three main contributions. Chapter 4 introduces the semi-supervised evidence fusion model with the application of MRI brain tumor segmentation. Chapter 5 describes the deep evidential segmentation framework to quantify segmentation uncertainty with 3D PET-CT lymphoma segmentation application. Chapter 6 presents the contextual discounting-based multimodal medical image fusion model with the application of multi-MRI brain tumor segmentation. Finally, we conclude our work and give some perspectives for future work in Chapter 7.

Part I

THEORETICAL BACKGROUND AND GENERAL CONTEXT

Chapter 1

Medical image segmentation

Contents

1.1 Introduction to medical image segmentation

Medical image segmentation is a critical step in radiotherapy planning, where organs at risk and tumors must be precisely located in images. The segmentation of medical images involves the extraction of regions of interest (ROI) from 2D/3D image data (e.g., pathological or optical imaging with color images, MRI, PET, and CT scans). The main goal of segmenting medical images is to identify areas of organs, cancer, and other abnormalities in the human body, for example, brain tumor, lymphomas, the interior of the human body (such as lungs, spinal canal, and vertebrae), and skin and cell lesions.

1.1.1 *Common medical imaging techniques*

Medical imaging is playing an increasingly central role in radiotherapy practice, thanks to the advances in imaging techniques. It influences the effectiveness of radiotherapy protocols, including the delineation of radiation target volumes and adjacent normal tissues, the design of the radiation dose distribution, the monitoring of treatment response, etc. A treatment that adopts advanced imaging technology to reduce uncertainties and assist decision-making during a course of treatment is referred to as image-guided radiotherapy. In this section, we introduce three commonly used medical imaging technologies: Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic resonance imaging (MRI).

COMPUTED TOMOGRAPHY (CT) CT is a gold standard image modality in radiation oncology, and its imaging devices are widely available at almost all cancer centers. CT images can provide anatomical information to show the geometric positions of the target tumor and adjacent organs at risk. CT imaging is based on the measurement of the X-rays attenuation between the source and the detector. The principle of CT imaging is briefly described in Figure 1.1. The use of CT in radiotherapy allows three-dimensional dose calculation and optimization.

POSITRON EMISSION TOMOGRAPHY (PET) PET is a functional imaging technique used in nuclear medicine that can measure tissue metabolic activity in vivo through an injected radioactive tracer. Complementary to CT, PET can provide critical functional information on target tissues or tumors for more precisely guarding the procedure of radiotherapy. The principle of PET imaging is briefly described in Figure 1.2. With different radioactive tracers, PET imaging makes it possible to monitor the different functional activities of a target tumor (e.g., metabolism, proliferation, and oxygen delivery) on a molecular scale. The most commonly used radioactive tracer in clinical oncology practice is fluorine-18 (F-18) fluorodeoxyglucose (FDG). PET scanning with FDG, i.e., FDG-PET, can highlight tumor tissues with high metabolic rate. It has been widely used for diagnosis, staging, and re-staging of most cancers, such as non-small cell lung cancer, esophageal carcinoma, or lymphomas, etc.

MAGNETIC RESONANCE IMAGING (MRI) MRI is a widely used technique to generate high-resolution and high-contrast medical images of soft tissues. With MRI, we can see inside the human body in detail by using magnets and radio to map the location of the water and then use this information to generate a detailed image. Figure 1.3 gives an introduction to the physics of MRI and its revelation of the brain. Different protocols or different MRI sequences defined by changing the repetition time (TR) and the echo time (TE) can be used to generate different modalities of MR images. For example, tissue can be characterized by two different relaxation times: T1 (longitudinal relaxation time) and T2 (transverse relaxation time).

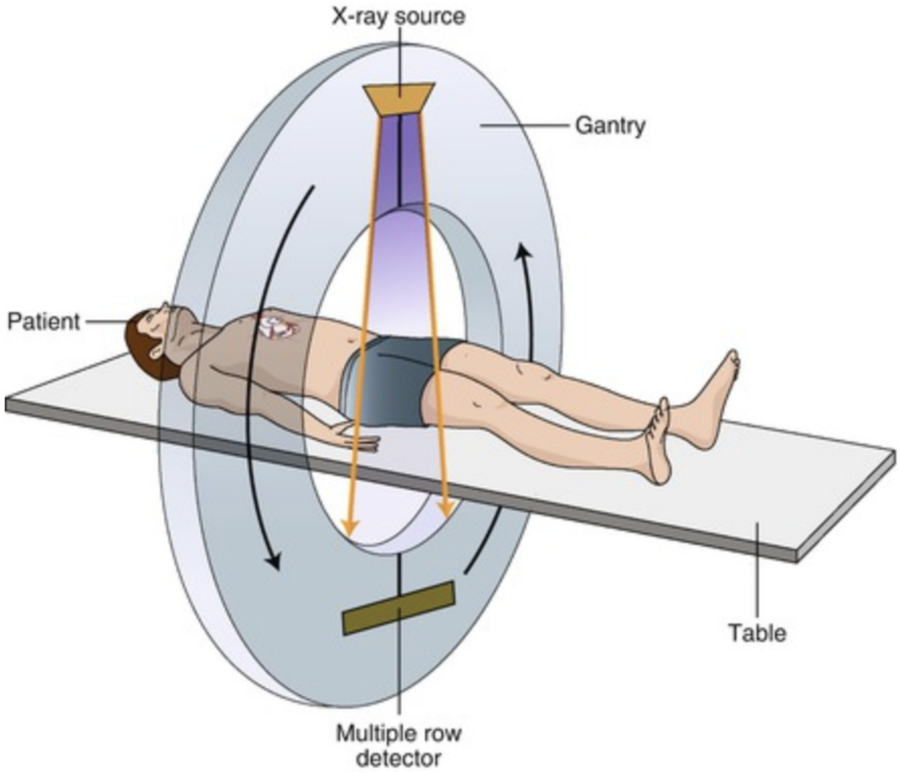


Figure 1.1: A depiction of a typical CT scanner [Pelberg, 2015]. The X-ray source and row detectors are housed within the gantry, which rotates around the patient (helical scanning technique). The cone-shaped X-ray beams emanate from the tube source on one side of the gantry, pass through the patient and terminate at the detectors on the other side of the gantry.

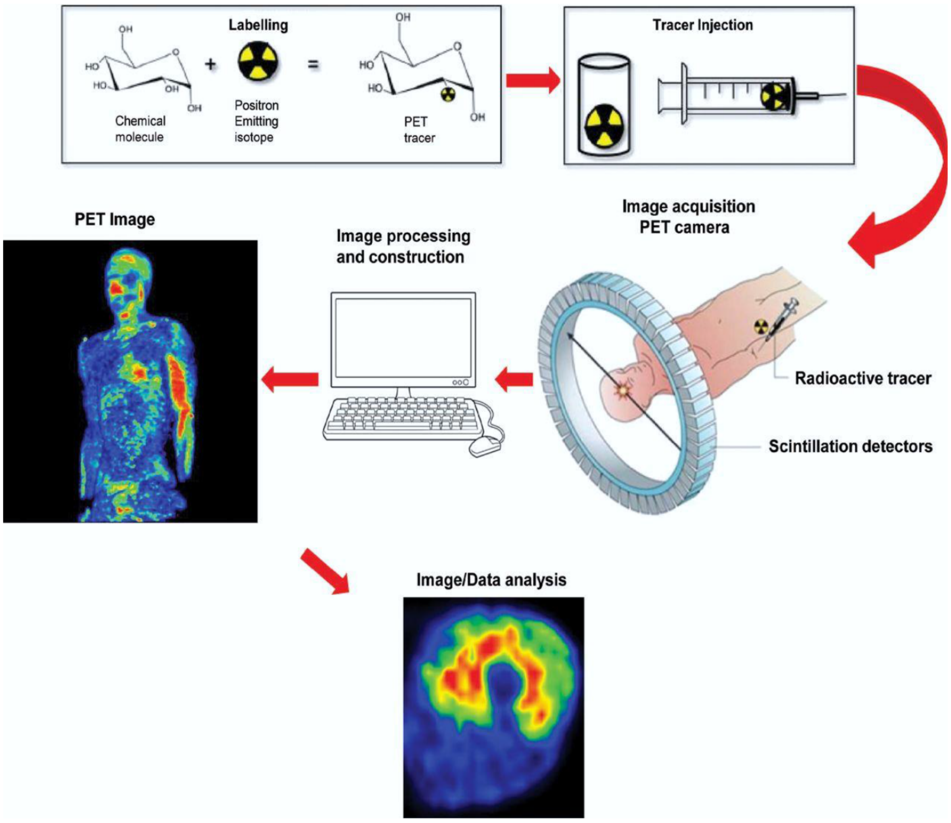


Figure 1.2: The fundamental principles of PET imaging [Rudroff et al., 2015]. PET tracer is first selected, and the information, either the tracer's kinetics within the tissues with dynamic imaging or the spatial distribution of the tracer with static imaging, is then detected by scintillation detectors of the PET scanner and is used to reconstruct images.

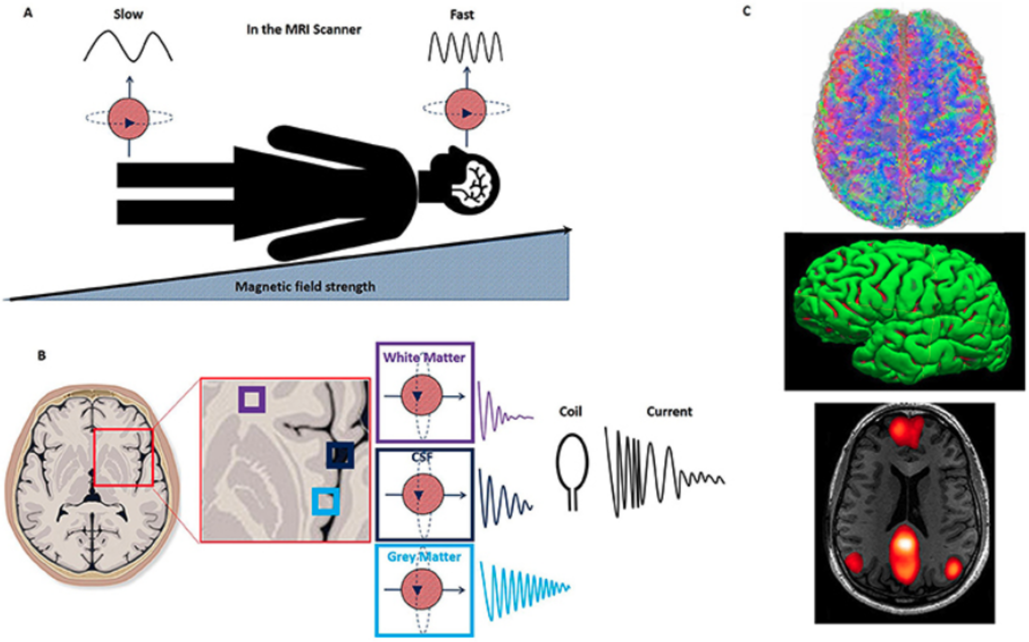


Figure 1.3: The physics of MRI and its revelation of the brain [Broadhouse, 2020]. (A) An example of an MRI scanner with a B1 field (magnetic field strength varies across the body), which increases across the body, from foot to head. Hydrogen protons in the head will then be spinning faster than those in the feet, allowing us to define the spatial position according to the frequencies. (B) Different tissues of a brain scan, such as white matter and gray matter (C) Examples of different brain images: the structural connection of the brain via white matter—the information (top right), the volume of gray matter regions of the brain (middle right), and the functional connection and communication of the brain (bottom right).

T1 determines the rate at which excited protons return to equilibrium in the direction of the main magnetic field, and T2 determines the rate at which excited protons reach equilibrium in the direction perpendicular to the main magnetic field or go out of phase with each other. The most common MRI sequences are T1-weighted and T2-weighted scans. T1-weighted images are produced by using the short TE and RT. This image weighting is useful for obtaining morphological information such as the cerebral cortex, fatty tissue, and focal liver lesions. In contrast, T2-weighted images are produced by using longer TE and TR. This image weighting is useful for detecting edema and inflammation, revealing white matter lesions, and assessing zonal anatomy in the prostate and uterus. Fluid-attenuated inversion recovery (FLAIR) is an MRI sequence with an inversion recovery set to null fluids. The FLAIR sequence is similar to a T2-weighted image except that the repetition time and TR times are very long. The FLAIR sequence is very sensitive to pathology and makes the differentiation between Cerebrospinal fluid (CSF) and an abnormality much easier. Radiotherapy is one of the five principal methods used in the clinical treatment of malignant tumors, the other methods being surgery, chemotherapy, immunotherapy, and hormonal therapy. Accurate tumor segmentation in medical images is a critical step for diverse objectives in clinical oncology, including reliable diagnosis and tumor staging, as well as solid radiotherapy planning. Thanks to the sustained advancement of medical imaging techniques, as well as progress made in medical image analysis, the effectiveness of radiotherapy for cancer treatment is being continuously improved.

1.1.2 *Significance of automatic medical image segmentation*

In the practical process of radiotherapy planning, the delineation of target tumor volumes is usually carried out manually by experienced clinicians with a computer interface. However, manual segmenting of medical images is time-consuming, labor-intensive, and operator-dependent. Moreover, the segmentation performance is sensitive to the operator's variability and may lead to imprecise and unreliable tumor contours. Recent advances in machine learning techniques, especially deep learning, make it possible to perform automatic segmentation with promising performance. In the rest of this chapter, we first briefly recall deep learning-based medical image segmentation methods in Section 1.2. Then we introduce three deep segmentation models that be used as the baseline models in this thesis in Section 1.3. Finally, in Section 1.4, we give a conclusion of existing deep learning-based segmentation methods and explain the advantages of using BFT with deep neural network in medical image segmentation.

1.2 DEEP LEARNING APPROACHES TO MEDICAL IMAGE SEGMENTATION

Early image segmentation methods used the information provided by the image itself, e.g., gray, textual, contrast, and histogram features, as well as segmenting ROI based on threshold [Batenburg et al., 2009], edge detection [Kimmel, 2003], graph partitioning [Onoma et al., 2014], clustering [Salvador et al., 2004], etc. More recently, researchers have been interested in hand-crafted features, e.g., Scale Invariant Feature Transform (SIFT) [Lowe, 1999], Features from Accelerated Segment Test (FAST) [Rosten et al., 2006], and Geometric hashing [Mian et al., 2006], as well as segmenting ROI using machine learning-based methods such as support vector machine (SVM) [Suykens et al., 1999], random forest (RF) [Strobl et al., 2007], logistic regression (LR) [Kleinbaum et al., 1996], etc. Those methods have attracted great interest for a while, but their accuracy cannot meet clinical application requirements because only low-level or middle-level features are considered.

Deep learning, a sub-field of machine learning concerned with neural networks, has achieved significant achievements in computer vision tasks by its ability to represent high-level semantic features [Krizhevsky et al., 2012; Bai et al., 2018], as well as in medical image segmentation domain [Ciresan et al., 2012; Seyedhosseini et al., 2013; Hariharan et al., 2015; Long et al., 2015; Ronneberger et al., 2015]. In [Ciresan et al., 2012], Ciresan et al. proposed a pixel-level classification network for electron microscopy (EM) image segmentation. This network won the competition of EM segmentation challenge at ISBI 2012 by a large margin. In [Seyedhosseini et al., 2013], Seyedhosseini et al. proposed a multiresolution contextual model called the cascaded hierarchical model (CHM). In [Hariharan et al., 2015], Hariharan et al. used hypercolumns as pixel descriptors to make the best use of semantic context and allow precise location. Long et al. [Long et al., 2015] were the first authors to show that a fully convolutional network (FCN) [Long et al., 2015] could be trained end-to-end for semantic segmentation, exceeding the state-of-the-art when the paper was published in 2015. UNet [Ronneberger et al., 2015], a successful modification and extension of FCN, has become the most popular model for medical image segmentation in recent years. Figure 1.4 shows an example framework of UNet with a down-sampling path and an up-sampling path. Based on UNet, research for deep learning-based medical image segmentation can be summarized in two major directions: the design of the segmentation model and the optimization (loss function).

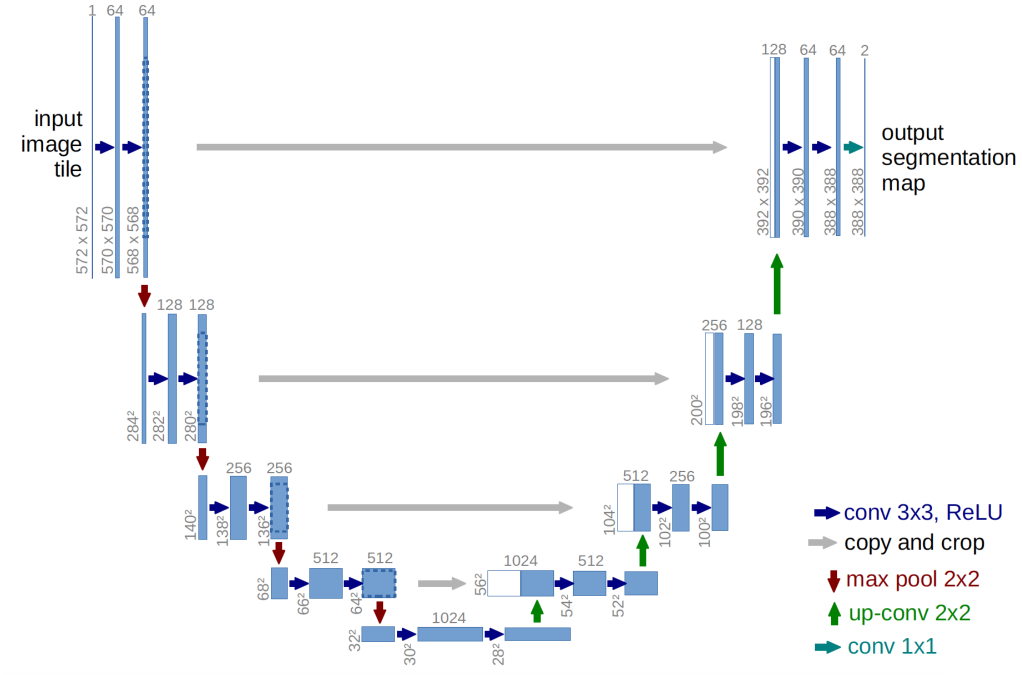


Figure 1.4: Example of an UNet architecture [Ronneberger et al., 2015]. Each blue box corresponds to feature maps (32 × 32 in the lowest resolution). The number of channels in the feature maps is denoted at the top of the box. The width and height of the convoluted feature maps are provided at the lower-left edge of the box. White boxes represent copied feature maps from the dotted box.

1.2.1 *Deep segmentation model*

Based on UNet, there has been many studies on deep segmentation networks with an encoder-decoder architecture (e.g., 3D UNet [Çiçek et al., 2016], Deep3D-UNet [Zhu et al., 2018], V-Net [Milletari et al., 2016], SegResNet [Myronenko, 2018]). Recently, some popular approaches, such as attention mechanism [Bahdanau et al., 2014] and transformer mechanism [Carion et al., 2020; Han et al., 2021] have achieved promising performance with deep neural networks (e.g., Attention UNet [Oktay et al., 2018; Trebing et al., 2021] and Transformer-UNet [Cao et al., 2021; Hatamizadeh et al., 2022b; Hatamizadeh et al., 2022a]), and have also been applied to medical image segmentation. Although the design of deep segmentation models with a research focus on the model structure has achieved great success, there are still some doubts about whether the so-called innovation of the network structure over the years is really just overfitting [Isensee et al., 2021], and some researchers suggest that more attention should be paid not only to the structure, but also to other aspects such as training and inference strategies. Thus, deep analysis of the segmentation performance, e.g., segmentation accuracy and uncertainty, is necessary to achieve a better segmentation performance.

1.2.2 *Model optimization*

Model optimization consists of designing loss functions relevant to the problems to be solved. Different from nature image segmentation, the segmented objects in a medical image are usually tiny compared with the background, which causes the unbalanced label distribution problem. The specific design of loss functions is one of the crucial ingredients in deep learning-based medical image segmentation. The existing loss functions for the deep segmentation model can be classified into four categories: distribution-based loss, region-based loss, boundary-based loss, and compound loss. Figure 1.5 shows the overview of loss functions for medical image segmentation. The distribution-based loss aims to minimize the dissimilarity between two distributions. Cross entropy (CE) loss is the fundamental function. Based on this, a lot of variants, such as Weighted CE loss [Ronneberger et al., 2015], Focal loss [Lin et al., 2017], distance penalized CE loss [Caliva et al., 2019], and CE with KL divergence [Zhou et al., 2022] are used for medical image segmentation. The region-based loss functions aim to minimize the mismatch or maximize the overlap regions between ground truth and predicted segmentation. The Dice loss [Milletari et al., 2016] is the most commonly used region-based loss function because it directly optimizes the Dice coefficient. Based on Dice loss, there are a lot of variants, e.g., IoU loss (also called Jaccard loss) [Rahman et al., 2016], Tversky loss [Salehi et al., 2017], Generalized Dice loss [Sudre et al., 2017],

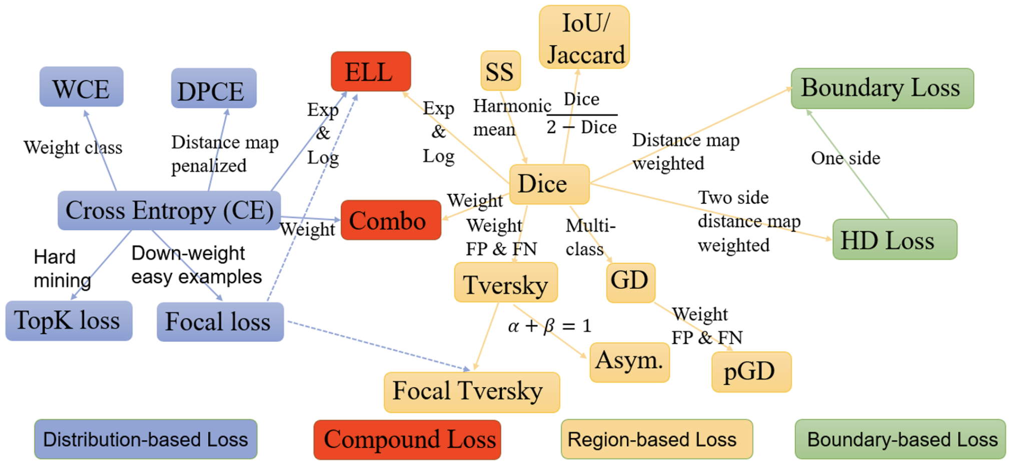


Figure 1.5: Overview of loss functions for medical image segmentation [Ma, 2020].

ContrastiveLoss [Chen et al., 2020], etc. The recently-proposed boundary-based loss functions aim to minimize a distance between ground truth and predicted segmentation. Representative examples are Boundary (BD) loss [Ma, 2020] and Hausdorff Distance (HD) loss [Karimi et al., 2019].

In addition to the above three categories of loss functions, compound loss functions are popular and efficient in medical image segmentation; they consist in combining different types of loss functions, i.e., the weighted sum between weighted CE and Dice loss. A general summary can be obtained from the existing literature that mildly imbalanced problems are well handled by Dice loss or generalized Dice loss [Sudre et al., 2017]. Highly imbalanced segmentation tasks are much more complicated and require more robust loss functions. More details about loss functions for deep segmentation can be found in [Ma, 2020].

1.2.3 *Evaluation criteria*

Dice score, Sensitivity, Precision, and Hausdorff distance (HD) are the most commonly used evaluation criteria to assess the quality of deep medical image segmentation methods; they are defined as follows:

$$\mathrm{Dice}(P,T) = \frac{2 \times TP}{FP + 2 \times TP + FN}, \tag{1.1}$$

$$\mathrm{Sensitivity}(P,T) = \frac{TP}{TP + FN}, \tag{1.2}$$

$$\mathrm{Precision}(P,T) = \frac{TP}{TP + FP}, \tag{1.3}$$

where P is the number of real positive voxels in the data, N is the number of real negative voxels in the data; TP, FP, and FN denote the numbers

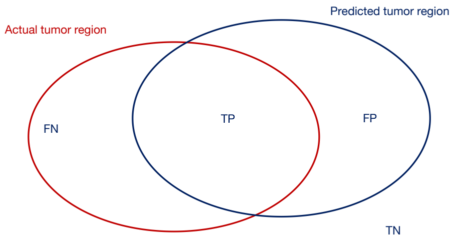


Figure 1.6: Interpretation of the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) used for the definition of evaluation criteria.

of true positive, false positive, and false negative voxels, respectively (See Figure 1.6), and

$$\text{HD} = \max\left(\max_{i\in S}\min_{j\in G} d(i,j), \max_{j\in G}\min_{i\in S} d(i,j)\right), \tag{1.4}$$

where G denotes the actual tumor region, S denotes the segmented tumor region, and d represents the distance of voxels between S and G.

The Dice score is a global measure of segmentation performance. Sensitivity is the proportion, among actual tumor voxels, of voxels correctly predicted as tumors. Precision is the proportion, among predicted tumor voxels, of voxels that actually belong to the tumor region; it is, thus, an estimate of the probability that the model is correct when it predicts that a voxel is in a tumor region. We note that neither sensitivity nor precision is global performance criteria. We can increase sensitivity by predicting the tumor class more often (at the expense of misclassifying a lot of background pixels), and we can increase precision by being very cautious and predicting the tumor class only when it has a high probability (at the expense of missing a lot of tumor voxels). These two criteria, thus, have to be considered jointly. Finally, we can also remark that a fourth criterion can also be defined: specificity, which is the proportion, among background voxels, of voxels, correctly predicted as background (i.e., $\text{TN}/(\text{TN}+\text{FP})$). However, as there are much more background voxels than tumor ones, this criterion is not informative in tumor segmentation applications (it is always very close to 1). The Hausdorff distance measures how far two subsets of a metric space are from each other with the definition in (1.4). More precisely, the Hausdorff distance is the greatest of all the distances from a point in one set to the closest point in the other set. Figure 1.7 shows an example of the calculation of the Hausdorff distance. A more comprehensive introduction to the evaluation criteria for medical image segmentation can be found in [Taha et al., 2015].

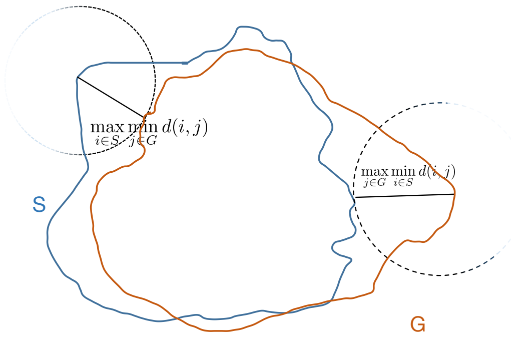

$\max_{i\in S}\min_{j\in G} d(i,j)$
$\max_{j\in G}\min_{i\in S} d(i,j)$
S
G

Figure 1.7: The calculation of the Hausdorff distance between the blue region S and the orange region G.

1.3 THREE BASELINE DEEP SEGMENTATION MODELS

In this section, we describe three deep segmentation networks: Residual-UNet [Kerfoot et al., 2018], MFNet [Chen et al., 2018], and No-New-UNet [Isensee et al., 2021], which will be used in the rest of the thesis as the baseline models.

1.3.1 *Residual-UNet*

Residual-UNet [Kerfoot et al., 2018] is a commonly used deep segmentation network with encoder and decoder layers defined using residual units. Data in the encoding path is down-sampled with convolution operation, and the decoding path is up-sampled using transpose convolution operation. Fig. 1.8 gives an example of Residual-UNet. Each Down-sampling layer (marked in blue) comprises convolution, normalization, dropout, and activation blocks, as well as a residual connection. Each Up-sampling layer (marked in green) comprises transpose convolution, normalization, dropout, activation blocks, and residual connection. The last layer (marked in yellow) is the bottom connection that does not down/up sample the data. The Residual-UNet will be used as the feature extraction module of the deep evidential segmentation model in Chapter 5 and the multimodal medical image segmentation model in Chapter 6.

1.3.2 *MFNet*

In [Chen et al., 2018], Chen et al. proposed a multi-fiber network (MFNet) whose framework is similar to UNet but can reduce by more than half the computation cost compared with UNet. It uses multi-fiber units instead of the residual units to construct the encoder-decoder module. As shown in Figure 1.9 (a), the conventional residual unit uses two $3 \times 3 \times 3$

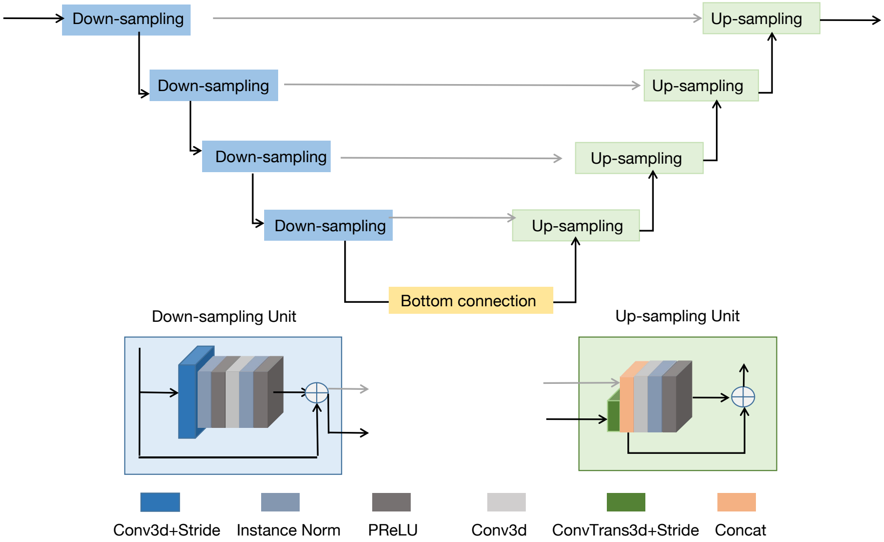


Figure 1.8: Framework of Residual-UNet. Each Down-sampling layer (marked in blue) comprises convolution, normalization, dropout, and activation blocks. Each Up-sampling layer (marked in green) comprises transpose convolution, normalization, dropout, and activation blocks. The last layer (marked in yellow) is the bottom connection which does not down/up sample the data.

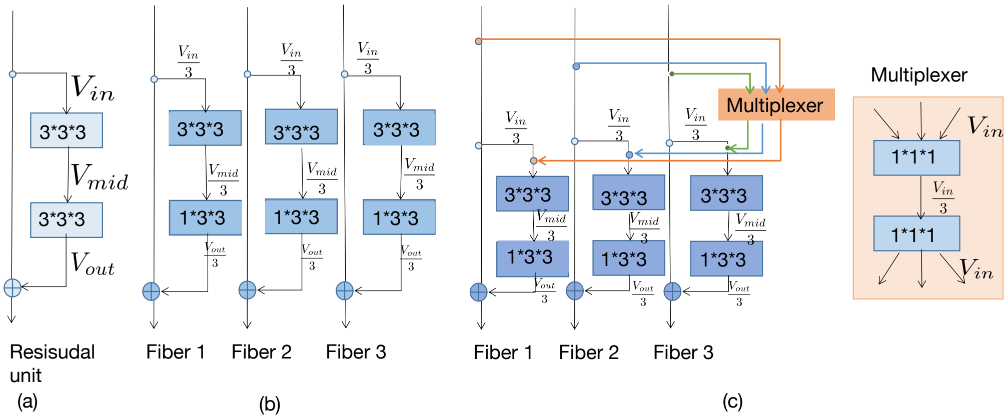


Figure 1.9: Illustration from residual unit to multi-fiber unit. (a) A single fiber with residual unit composed of two $3 \times 3 \times 3$ convolution layers [He et al., 2016]. (b) Multi-fiber (MF) units [Chen et al., 2018], 3 fibers for example. (c) Multi-fiber units with a multiplexer for transferring information across separated fibers [Chen et al., 2018]. Let V_{in}, V_{mid}, and V_{out} denote the number of input channels, middle channels, and output channels, respectively. The total number of connections for the residual unit and multi-fiber units is $V_{in} \times V_{mid} + V_{mid} \times V_{out}$ and $(V_{in} \times V_{mid} + V_{mid} \times V_{out})/M$, respectively.

convolutional layers to learn features, which is straightforward but computationally expensive. The multi-fiber unit is illustrated in Figure 1.9 (b). It slices the residual unit into M parallel and separated paths (called fibers here). The connection number is then reduced by the slicing parameter M. However, the slicing operation isolates each path from the others and blocks information flow across them, resulting in limited learning capacity for information representation. Thus, the authors propose a multiplexer layer (see Figure 1.9 (c)) that acts as a router that operates across fibers. The multiplexer layer first gathers features from all fibers using a $1 \times 1 \times 1$ convolution layer and then redirects them to specific fibers using the following $1 \times 1 \times 1$ convolution layer. The MFNet unit will be used as the feature extraction module of the semi-supervised medical image segmentation model in Chapter 4.

1.3.3 *No-New-UNet (nnUNet)*

nnUNet [Isensee et al., 2021] is the first segmentation model designed as a segmentation pipeline for any given dataset and is a plug-and-play tool for state-of-the-art biomedical segmentation. Instead of designing a new encoder-decoder structure, the authors focus on studying a recipe that systematizes the configuration process on a task-agnostic level and drastically reduces the search space for empirical design choices by a set of fixed parameters, interdependent rules, and empirical decisions. Due to its modular structure, new architectures or algorithms can easily be integrated into nnUNet. Figure 1.10 is the proposed automated configuration by nnUNet. The nnUNet will be used as another feature extraction module of the multimodal medical image segmentation model in Chapter 6.

1.4 CONCLUSION

The research reviewed in this chapter focuses on improving the accuracy of segmentation performance under the assumption of adequate and perfect input information and accurate and appropriate prior knowledge. However, in reality, especially in the medical image segmentation domain, both the input information and prior knowledge are imperfect and contain a degree of uncertainty. Figure 1.11a illustrates uncertain information taking a brain tumor segmentation task as an example. Let X be the type of tumor of a voxel, and $\Omega = \{ED, ET, NRC, Others\}$, corresponding to the possibilities: edema, enhancing tumor, necrotic core, and others. Let us assume that a specialist provides the information $X \in \{ED, ET\}$, but there is a probability $p = 0.1$ that the information is unreliable. How to represent this situation by a probability function is a challenging problem. Another situation is when we have multiple information sources tainted with uncertainty, as illustrated in Figure 1.11b; how can we model that kind of uncertainty and

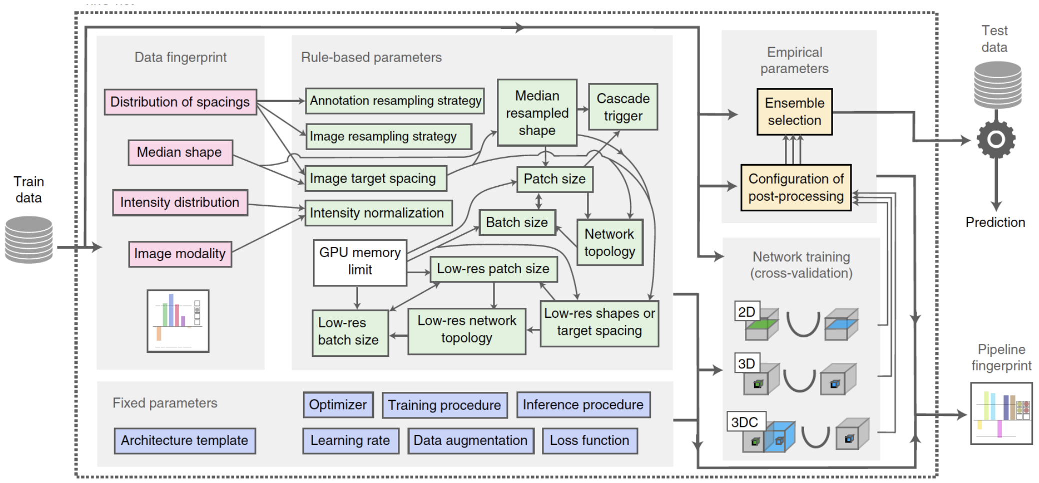


Figure 1.10: Proposed automated configuration by nnUNet [Isensee et al., 2021]. Dataset properties are summarized in a "dataset fingerprint". A set of heuristic rules operates on this fingerprint to infer the data-dependent hyperparameters of the pipeline. These are completed by blueprint parameters and the data-independent design choices to form "pipeline fingerprints". Three architectures are trained based on these pipeline fingerprints in a 5-fold cross-validation. Finally, nnU-Net automatically selects the optimal ensemble of these architectures and performs post-processing if required.

fuse the evidence? Furthermore, if the information sources are in conflict and contain uncertainty as well, i.e., Figure 1.11c, it is difficult to represent and summarize that information by probabilistic models. Thanks to BFT, these challenges can be addressed by designing new frameworks for modeling, reasoning, and fusing imperfect (uncertain, imprecise) information. In the next chapter, we will give a brief introduction to BFT.

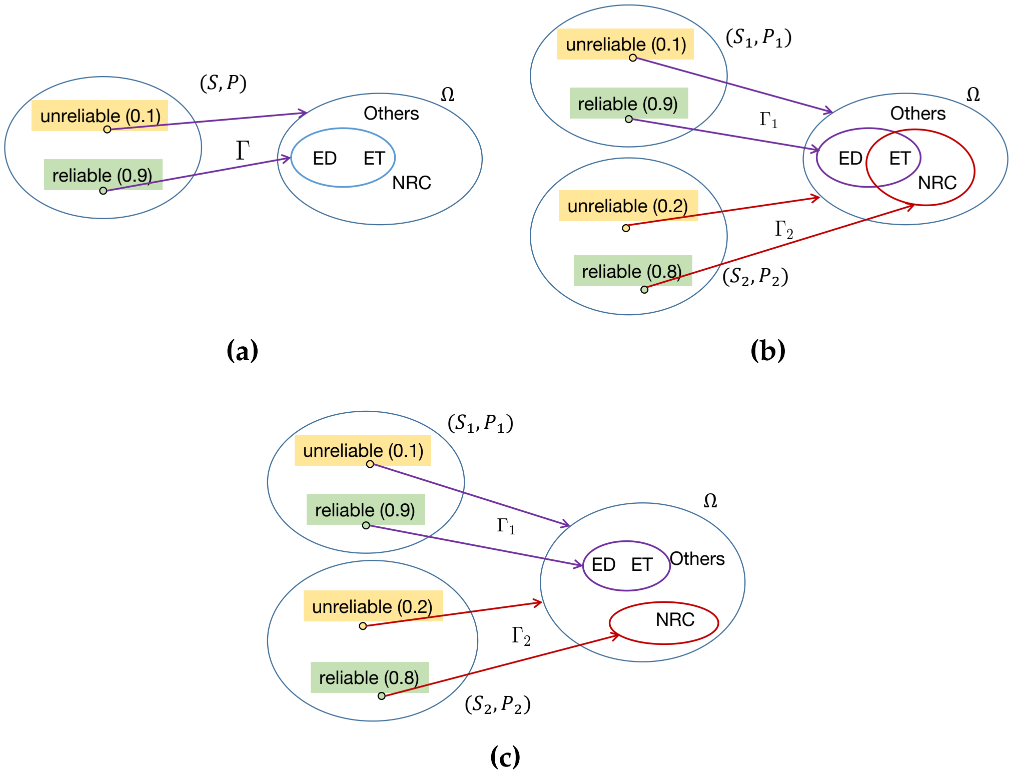


Figure 1.11: (a) Example of a segmentation task with uncertain information, (b) Example of a segmentation task with multiple sources of information, (c) Example of a segmentation task with conflicting sources of information.

Chapter 2

Fundamentals of belief function theory

Contents

BFT is a generalization of Bayesian theory, but it is more flexible than the Bayesian approach and suitable under weaker conditions [Sun et al., 2018], i.e., imperfect (uncertain, imprecise, partial) information. Figure 2.1 shows the difference between a probabilistic model and a BFT model when applied to a three-class classification task ($\Omega = \{a, b, c\}$). For input x, the probabilistic model outputs the probability that x belongs to classes a, b, and c as 0.4, 0.5, and 0.1, respectively. In contrast, the BFT model can represent degrees of belief that x belongs specifically to any subset of Ω, e.g., $\{a, b\}$, $\{b, c\}$. Compared with the probabilistic model, the BFT model has more degrees of freedom to represent its uncertainty directly, as it shares a unit mass of belief among all subsets of Ω.

In the past decades, BFT has generated considerable interest and has had great success in diverse fields, including uncertain reasoning [Smets, 1990; Yager, 1987; Dubois et al., 1988; Denœux, 2008], classification [Denœux, 1995; Denœux, 2000] and clustering [Denœux et al., 2004; Masson et al., 2009], etc. It was first originated by Dempster [Dempster, 1967] in the context of statistical inference in 1968 and was later formalized by Shafer [Shafer, 1976] as a theory of evidence in 1976. In 1986, Dubois and Prade proposed an approach to the computerized

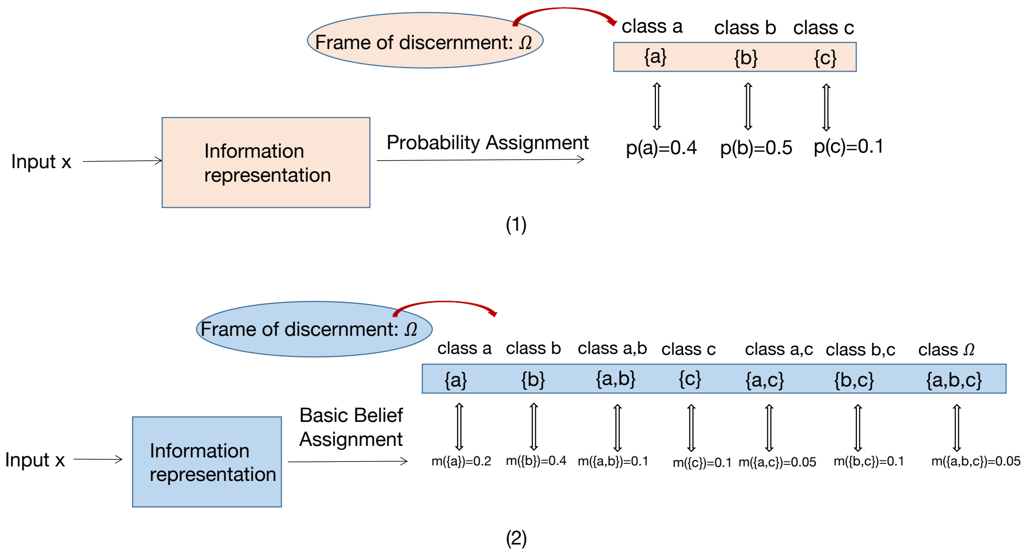


Figure 2.1: An example of three class assignments: (1) Probabilistic model and (2) BFT model. In contrast with the probabilistic model, the BFT model can quantify uncertainty and assign it to the focal set $\{a, b\}$,$\{a, c\}$,$\{b, c\}$ and $\{a, b, c\}$ to represent its uncertainty or ignorance (m here is the evidence (mass function) about a variable ω taking values in Ω, which will be introduced in Section 2.1).

processing of uncertainty [Dubois et al., 2012]. In 1978, Yager proposed a new combination rule of the belief function framework [Yager, 1987]. In 1990, BFT was further popularized and developed by Smets [Smets, 1990] as the 'Transferable Belief Model' with the pignistic transformation for decision making. Since then, there has been many developments. More detailed information about the development of BFT in 40 years can be found in [Denœux, 2016].

In this chapter, we first introduce the basic notions of BFT, which includes evidence representation (i.e., mass functions, belief, plausibility and contour function, as well as simple mass function) in Section 2.1. Second, we introduce Dempster's rule in Section 2.2 to explain fusion operations of multiple sources of evidence. Third, we introduce the discounting operation for unreliable sources in Section 2.3. Fourth, we introduce some commonly used decision-making methods in Section 2.4. Finally, in Section 2.5, we give a summary of basic belief assignment methods used to generate mass functions for medical image segmentation.

2.1 REPRESENTATION OF EVIDENCE

Let $\Omega = \{\omega_1, \omega_2, ..., \omega_C\}$ be a finite set of all possible hypotheses about some problem, called the frame of discernment. Evidence about a variable

ω taking values in Ω can be represented by a mass function m, from the power set 2^Ω to $[0, 1]$, such that

$$\sum_{A \subseteq \Omega} m(A) = 1, \tag{2.1a}$$

$$m(\emptyset) = 0. \tag{2.1b}$$

Mapping m can also be called **basic belief assignment** (BBA). Each subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal set of m. The mass $m(\Omega)$ represents the degree of ignorance about the problem. If all focal sets are singletons, then m is said to be Bayesian and it is equivalent to a probability distribution.

The information provided by a mass function m can also be represented by a **belief function** Bel or a **plausibility function** Pl from 2^Ω to $[0, 1]$ defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2.2}$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}), \tag{2.3}$$

for all $A \subseteq \Omega$, where $\bar{A}$ denotes the complement of A. The quantity $Bel(A)$ can be interpreted as a degree of support to A, while $Pl(A)$ can be interpreted as a measure of lack of support given to the complement of A. The **contour function** pl associated to m is the function that maps each element ω of Ω to its plausibility:

$$pl(\omega) = Pl(\{\omega\}), \quad \forall \omega \in \Omega. \tag{2.4}$$

A mass function m is said to be *simple* if it can be obtained by discounting a logical mass function; it thus has the following form:

$$m(A) = s, \quad m(\Omega) = 1 - s, \tag{2.5}$$

for some $A \subset \Omega$ such that $A \neq \emptyset$ and some $s \in [0, 1]$, called the *degree of support* in A. The quantity $w = -\ln(1 - s)$ is called the *weight of evidence* associated to m [Shafer, 1976, page 77]. In Chapter 5, a **simple mass function** with focal set A and weight of evidence w will be denoted as A^w.

2.2 DEMPSTER'S RULE

In BFT, the belief about a certain question is elaborated by aggregating different belief functions over the same frame of discernment. Given two

Table 2.1: Example of Dempster's rule for evidence fusion.

m_2 m_1	$\{a\}$, 0.2	$\{b\}$, 0.3	$\{a,b\}$, 0.1	$\{c\}$, 0.1	$\{a,c\}$, 0	$\{b,c\}$,0.2	$\{a,b,c\}$, 0.1
$\{a\}$, 0.3	$\{a\}$, 0.06	$\emptyset$, 0.09	$\{a\}$, 0.03	$\emptyset$, 0.03	$\{a\}$, 0	$\emptyset$, 0.06	$\{a\}$, 0.03
$\{b\}$, 0.3	$\emptyset$, 0.06	$\{b\}$, 0.09	$\{b\}$, 0.03	$\emptyset$, 0.03	$\emptyset$,0	$\{b\}$, 0.06	$\{b\}$, 0.03
$\{a,b\}$, 0.1	$\{a\}$, 0.02	$\{b\}$, 0.03	$\{a,b\}$, 0.01	$\emptyset$, 0.01	$\{a\}$, 0	$\{b\}$, 0.02,	$\{a,b\}$, 0.01
$\{c\}$, 0	$\emptyset$, 0	$\emptyset$, 0	$\emptyset$, 0	$\{c\}$, 0	$\{c\}$, 0	$\{c\}$, 0	$\{c\}$, 0
$\{a,c\}$, 0.1	$\{a\}$, 0.02	$\emptyset$, 0.03	$\{a\}$, 0.01	$\{c\}$, 0.01	$\{a,c\}$, 0	$\{c\}$, 0.02	$\{a,c\}$, 0.01
$\{b,c\}$, 0.1	$\emptyset$, 0.02	$\{b\}$, 0.03	$\{b\}$, 0.01	$\{c\}$, 0.01	$\{c\}$, 0	$\{b,c\}$, 0.02	$\{b,c\}$, 0.01
$\{a,b,c\}$, 0.1	$\{a\}$, 0.02	$\{b\}$, 0.03	$\{a,b\}$, 0.01	$\{c\}$, 0.01	$\{a,c\}$, 0	$\{b,c\}$, 0.02	$\{a,b,c\}$, 0.01

mass functions m_1 and m_2 derived from two independent items of evidence, the final belief that supports A can be obtained by combining m_1 and m_2 with Dempster's rule [Shafer, 1976] defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap D = A} m_1(B) m_2(D), \tag{2.6}$$

for all $A \subseteq \Omega, A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. The coefficient κ is the degree of conflict between m_1 and m_2, it is defined as

$$\kappa = \sum_{B \cap D = \emptyset} m_1(B) m_2(D). \tag{2.7}$$

Mass functions m_1 and m_2 can be combined if and only if $\kappa < 1$. The mass function $m_1 \oplus m_2$ is called the orthogonal sum of m_1 and m_2. Let pl_1, pl_2 and pl_{12} denote the contour functions associated with, respectively, m_1, m_2 and $m_1 \oplus m_2$. The following equation holds:

$$pl_{12} = \frac{pl_1 pl_2}{1-\kappa}. \tag{2.8}$$

Table 2.1 shows an example of two source evidence fusion by using Dempster's rule for $A = \{a\},\{b\},\{a,b\},\{c\},\{a,c\},\{b,c\}$, and $\{a,b,c\}$. The degree of conflict is $\kappa = 0.09 + 0.03 + 0.06 + 0.06 + 0.03 + 0.01 + 0.03 + 0.02 = 0.33$. The combined mass function is

$$(m_1 \oplus m_2)(\{a\}) = (0.06 + 0.03 * 2 + 0.02 * 3 + 0.01)/(1 - 0.33) = 19/67,$$

$$(m_1 \oplus m_2)(\{b\}) = (0.09 + 0.03 * 5 + 0.06 + 0.02 + 0.01)/0.67 = 33/67,$$

$$(m_1 \oplus m_2)(\{a,b\}) = (0.01 * 3)/0.67 = 3/67,$$

$$(m_1 \oplus m_2)(\{c\}) = (0.01 * 3 + 0.02)/0.67 = 5/67,$$

$$(m_1 \oplus m_2)(\{a,c\}) = 0.01/0.67 = 1/67,$$

$$(m_1 \oplus m_2)(\{b,c\}) = (0.02 * 2 + 0.01)/0.67 = 5/67,$$

$$(m_1 \oplus m_2)(\{a,b,c\}) = 0.01/0.67 = 1/67.$$

2.3 DISCOUNTING

In (2.7), if m_1 and m_2 are logically contradictory, we cannot use Dempster's rule to combine them. Discounting strategies can be used to combine highly conflicting evidence [Shafer, 1976; Mercier et al., 2008; Denœux et al., 2019]. Let m be a mass function on Ω and β a reliability coefficient in $[0,1]$. The *discounting* operation [Shafer, 1976] with the discount rate $1-\beta$ transforms m into a weaker, less informative mass function defined as follows:

$$^{\beta}m = \beta\, m + (1-\beta)\, m_?, \tag{2.9}$$

where $m_?$ is the vacuous mass function defined by $m_?(\Omega)=1$, and coefficient β is the degree of belief that the source mass function m is reliable [Smets et al., 1994]. When $\beta = 1$, we accept the mass function m provided by the source and take it as a description of our knowledge; when $\beta = 0$, we reject it, and we are left with the vacuous mass function $m_?$. In this paper, we focus on the situation when $\beta \in [0,1]$ and combine uncertain evidence with partial reliability using Dempster's rule. Mercier et al. extended the discounting operation with *contextual discounting* in the BFT framework to refine the modeling of sensor reliability, allowing us to use more detailed information regarding the reliability of the source in different contexts [Mercier et al., 2008], i.e., conditionally on different hypotheses regarding the variable of interest. More details about contextual discounting will be introduced in Chapter 6.

2.4 DECISION-MAKING (DM)

After combining all the available evidence in the form of a mass function, it is necessary to make a decision. In this section, we introduce some classical BFT-based decision-making methods.

UPPER AND LOWER EXPECTED UTILITIES Let u be a utility function. The lower and upper expectations of u with respect to m are defined, respectively, as the averages of the minima and the maxima of u within each focal set of m:

$$\underline{E}_m(u) = \sum_{A\subseteq\Omega} m(A) \min_{\omega\in A} u(\omega), \tag{2.10a}$$

$$\overline{E}_m(u) = \sum_{A\subseteq\Omega} m(A) \max_{\omega\in A} u(\omega). \tag{2.10b}$$

When m is Bayesian, $\underline{E}_m(u) = \overline{E}_m(u)$. If m is logical with focal set A, then $\underline{E}_m(u)$ and $\overline{E}_m(u)$ are, respectively, the minimum and maximum of u in

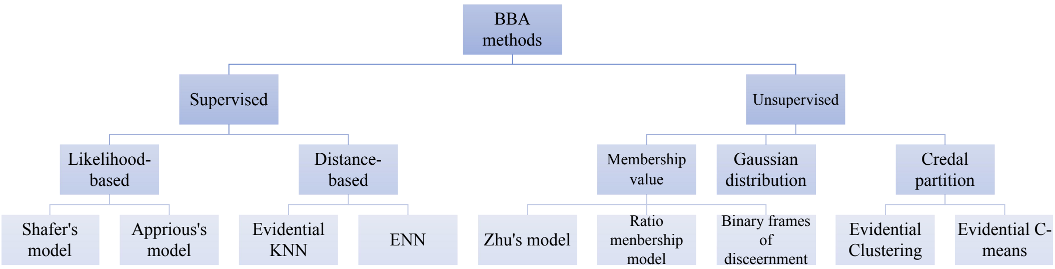


Figure 2.2: Overview of Basic belief assignment (BBA) methods.

A. The lower or upper expectations can be chosen for the final decision according to the given task and decision-maker's attitude.

PIGNISTIC CRITERION In 1990, Smets proposed a pignistic transformation [Smets, 1990] that distributes each mass of belief distributed equally among the elements of Ω. The pignistic probability distribution is defined as

$$\mathrm{BetP}(\omega) = \sum_{\omega \in A} \frac{m(A)}{\mid A \mid}, \quad \forall \omega \in \Omega, \tag{2.11}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$.

Besides the above methods, there are various decision-making methods proposed for BFT, such as Generalized OWA criterion [Yager, 1992], Generalized minimax regret [Yager, 2004], Generalized divergence [Xiao et al., 2022], etc. More details about decision-making with BFT can be found in the review paper [Denœux, 2019a].

2.5 BASIC BELIEF ASSIGNMENT (BBA) METHODS TO GENERATE MASS FUNCTIONS

To model segmentation uncertainty, the first step is to generate mass functions. In this section, we introduce the main basic belief assignment methods applied to medical image segmentation. Figure 2.2 is an overview of basic belief assignment methods. In general, those methods can be separated into supervised and unsupervised methods according to whether annotations are used to optimize the parameters of basic belief assignment models or not.

2.5.1 *Supervised basic belief assignment methods*

Supervised basic belief assignment methods can be classified into two categories. One is the likelihood-based methods, such as Shafer's model [Shafer, 1976] and Appriou's model [Appriou, 1999; Appriou, 2005]. The other category is composed of distance-based methods, such as the evidential KNN rule [Denœux, 1995], the evidential neural network classifier [Denœux, 2000] and Radial basis function networks [Denœux, 2019a]. It should be noted that the distance-based methods can easily be merged with popular deep segmentation models and have shown promising results [Tong et al., 2021b; Huang et al., 2021a; Huang et al., 2022b].

2.5.1.1 *Likelihood-based methods*

SHAFER'S MODEL In [Shafer, 1976], Shafer proposed a likelihood-based evidential model to calculate mass functions. Assuming that conditional density functions $f(x \mid \omega_c)$ are known, then the conditional likelihood associated with the pattern X is defined by $\ell(\omega_c \mid x) = f(x \mid \omega_c)$. The mass functions are defined according to the knowledge of all hypotheses $\omega_1, \dots, \omega_C$. Firstly, the plausibility of a simple hypothesis ω_c is proportional to its likelihood. The plausibility is, thus, given by

$$Pl(\{\omega_c\}) = \hbar \cdot \ell(\omega_c \mid x), \quad \forall \omega_c \in \Omega, \tag{2.12}$$

where $\hbar$ is a normalization factor with $\hbar = 1/\max_{\omega \in \Omega} \ell(\omega|x)$. The plausibility of a set A is, thus, given by

$$Pl(A) = \hbar \cdot \max_{\omega_c \in A} \ell(\omega_c \mid x). \tag{2.13}$$

APPRIOU'S MODEL Appriou [Appriou, 1999; Appriou, 2005] also proposed two likelihood-based models to calculate mass functions with the frame of discernment $\Omega = \{\omega_c, \neg\omega_c\}$. For the first model, the mass functions are defined by

$$\begin{aligned} m(\{\omega_c\}) &= 0, && (2.14a)\\ m(\{\neg\omega_c\}) &= \alpha_c(1 - \hbar \cdot \ell(\omega_c \mid x)), && (2.14b)\\ m(\Omega) &= 1 - \alpha_c(1 - \hbar \cdot \ell(\omega_c \mid x)), && (2.14c) \end{aligned}$$

where α_c is a reliability factor depending on the hypothesis ω_c and on the source information. The second model is defined as

$$\begin{aligned} m(\{\omega_c\}) &= \alpha_c \cdot \hbar \cdot \ell(\omega_c \mid x)/(1 + \hbar \cdot \ell(\omega_c \mid x)), && (2.15a)\\ m(\{\neg\omega_c\}) &= \alpha_c/(1 + \hbar \cdot \ell(\omega_c \mid x)), && (2.15b)\\ m(\Omega) &= 1 - \alpha_c. && (2.15c) \end{aligned}$$

2.5.1.2 *Distance-based methods*

EVIDENTIAL KNN (EKNN) RULE In [Denœux, 1995], Denœux proposed a distance-based KNN classifier for classification tasks. Let $N_K(x)$ denote the set of the K nearest neighbors of x in learning set. Each $x_i \in N_K(x)$

is considered as a piece of evidence regarding the class label of x. The strength of evidence decreases with the distance between x and x_i. The evidence of (x_i, y_i) support class c is represented by

$$\begin{aligned} m_i(\{\omega_c\}) &= \varphi_c(d_i) y_{ic}, \quad 1 \leqslant c \leqslant C, && (2.16a) \\ m_i(\Omega) &= 1 - \varphi_c(d_i), && (2.16b) \end{aligned}$$

where d_i is the distance between x and x_i, which can be the Euclidean or other distance function; and $y_{ic} = 1$ if $y_i = \omega_c$ and $y_{ic} = 0$ otherwise. Function φ_c is defined as

$$\varphi_c(d) = \alpha \exp(-\gamma d^2), \tag{2.17}$$

where α and γ are two tuning parameters. The evidence of the K nearest neighbors of x is fused by Dempster's rule:

$$m = \bigoplus_{x_i \in N_K(x)} m_i. \tag{2.18}$$

The final decision is made according to maximum plausibility. The detailed optimization of these parameters is described in [Zouhal et al., 1998]. Based on this first work, Denœux et al. proposed the contextual discounting evidential KNN rule [Denœux et al., 2019] with partially supervised learning to address the annotation limitation problem.

EVIDENTIAL NEURAL NETWORK (ENN) The success of machine learning encouraged the exploration of applying belief function theory with learning methods. In [Denœux, 2000], Denœux proposed an ENN classifier in which mass functions are computed based on distances to prototypes.

The ENN classifier is composed of an input layer of H neurons, two hidden layers, and an output layer. The first input layer is composed of I units, whose weights vectors are prototypes $p_1, \ldots, p_I$ in input space. The activation of unit i in the prototype layer is

$$s_i = \alpha_i \exp(-\gamma_i d_i^2), \tag{2.19}$$

where $d_i = |x - p_i|$ is the Euclidean distance between input vector x and prototype p_i, $\gamma_i > 0$ is a scale parameter, and $\alpha_i \in [0, 1]$ is an additional parameter. The second hidden layer computes mass functions m_i representing the evidence of each prototype p_i, using the following equations:

$$\begin{aligned} m_i(\{\omega_c\}) &= u_{ic} s_i, \quad c = 1, \ldots, C && (2.20a) \\ m_i(\Omega) &= 1 - s_i, && (2.20b) \end{aligned}$$

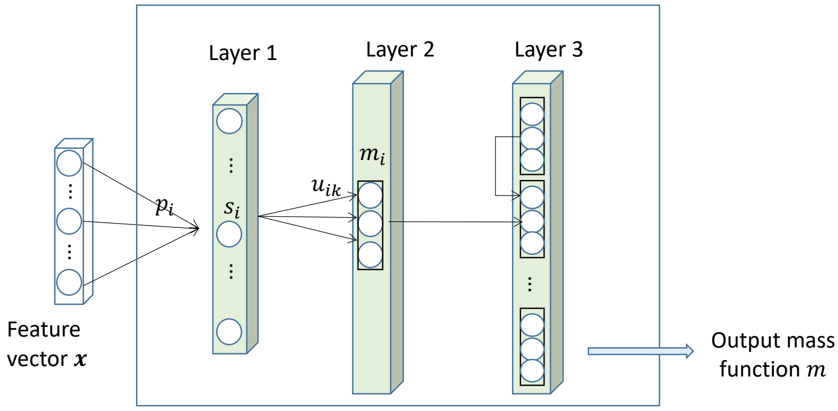


Figure 2.3: Evidential neural network.

where u_{ic} is the membership degree of prototype i to class ω_c, and $\sum_{c=1}^{C} u_{ic} = 1$. Finally, the third layer combines the I mass functions $m_1, \dots, m_I$ using Dempster's rule. The output mass function $m = \bigoplus_{i=1}^{I} m_i$ is a discounted Bayesian mass function that summarizes the evidence of the I prototypes.

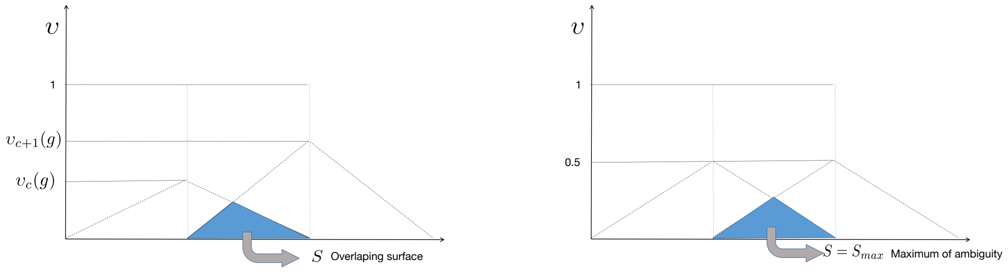
Let θ denote the vector of all network parameters, composed of the I prototypes $\mathbf{p}_i$, their parameters γ_i and α_i, and their membership degrees u_{ic}, $c = 1, \dots, C$. In [Denœux, 2000], it was proposed to learn these parameters by minimizing the regularized sum-of-squares loss function

$$L_{SS}(\theta) = \sum_{n=1}^{N} \sum_{c=1}^{C} (p_{nc} - y_{nc})^2 + \lambda \sum_{i=1}^{I} \alpha_i, \tag{2.21}$$

where p_{nc} is the pignistic probability of class ω_c for instance n, N is the number of training instances, and $y_{nc} = 1$ if the true class of instance n is ω_c, and $y_{nc} = 0$ otherwise. The second term on the right-hand side of (2.21) is a regularization term, and λ is a hyperparameter that can be tuned by cross-validation.

2.5.2 *Unsupervised basic belief assignment methods*

The goal of unsupervised basic belief assignment methods is to generate mass functions without any label information. In earlier basic belief assignment studies, Fuzzy C-means (FCM) [Dunn, 1973] was the most popular algorithm used to generate membership values (MVs). Based on MVs, the authors can obtain mass functions according to some domain knowledge, e.g., threshold [Zhu et al., 2002], or user-specific parameters [Ghasemi et al., 2012]. The sigmoid and one-sided Gaussian function can also be used to generate MVs [Safranek et al., 1990]. The notion of credal partition [Denœux et al., 2004], an extension of fuzzy partition, enables us to generate mass functions directly [Masson et al., 2008]. Besides these two



(a) **(b)**

Figure 2.4: (a) Construction of triangular membership functions and (b) maximum ambiguity case: $\upsilon_c(g) = \upsilon_{c+1}(g) = 0.5$.

popular basic belief assignment methods, mass functions can also be generated from the Gaussian distribution of the input to the cluster center [Chen et al., 2012].

2.5.2.1 *MV-based methods*

FCM Considering that there are some BFT-based methods that use FCM to generate MVs, we briefly summarize FCM here to offer a basic view for readers. With FCM, any $\mathbf{x}$ has a set of coefficients $w_k(x)$ representing the degree of membership in the kth cluster. The centroid of a cluster is the mean of all points, weighted by the m-th power of their membership degree,

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}, \tag{2.22}$$

where m is the hyper-parameter that controls how fuzzy the cluster will be. The higher it is, the fuzzier. Given a finite set of data, the FCM algorithm returns a list of cluster centers $P = \{c_1, \ldots, c_C\}$ and a partition matrix $W = (w_{ij}), i = 1, \ldots, N, j = 1, \ldots, C$,

$$w_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_c\|} \right)^{\frac{2}{m-1}}}, \tag{2.23}$$

where w_{ij}, is the degree of membership of $\mathbf{x}_i$ to cluster $\mathbf{c}_j$. The objective function is defined as

$$\arg\max_P \sum_{i=1}^{N} \sum_{j=1}^{C} w_{ij}^m \left\|\mathbf{x}_i - \mathbf{c}_j\right\|. \tag{2.24}$$

ZHU'S MODEL In [Zhu et al., 2002], Zhu et al. proposed a method to determine mass functions using FCM and neighborhood information. The mass assigned to a simple hypothesis $\{\omega_c\}$ is directly obtained from the filtered membership functions $\upsilon_c(g)$ of the gray level $g(x,y)$ to cluster c as $m(\{\omega_c\}) = \upsilon_c(g)$. For a given gray level, the piece of evidence of belonging to the cluster c is, thus, directly given by its degree of membership to the same cluster. If there is a high ambiguity in assigning a gray level $g(x,y)$ to cluster c or $c+1$, that is, $|\upsilon_c(g) - \upsilon_{c+1}(g)| < \varepsilon$, where ε is a thresholding value, then a double hypothesis is formed. The value of the threshold ε is chosen depending on the application. The authors suggested fixing ε at 0.1. Once the double hypotheses are formed, their associated mass is calculated according to the following formula:

$$m(\{\omega_c, \omega_{c+1}\}) = \frac{S[\upsilon_c(g), \upsilon_{c+1}(g)]}{2S_{max}}, \tag{2.25}$$

where S represents the surface of a triangle and S_{max} is the maximum of ambiguity. The surface of such a triangle depends both on the degrees of the membership functions of $g(x,y)$ to clusters c and $c+1$ and on the conflicts between these MVs. Figure 2.4a shows how the triangle is constructed and how the mass of double hypotheses $\{\omega_c, \omega_{c+1}\}$ is derived from the surface of the triangle. The vertical axis of Figure 2.4a represents the MVs. The surfaces of the two dotted triangles define two so-called triangular membership functions corresponding to classes c and $c+1$. The two triangles are isosceles and have the same length for their bases. The heights of the triangles are equal to $m(\{\omega_c\})$ and $m(\{\omega_{c+1}\})$, respectively. The overlapping surface S of the two triangles represents the MV to the double hypothesis $\{\omega_c, \omega_{c+1}\}$. Therefore, the mass value attributed to the double hypothesis $\{\omega_c, \omega_{c+1}\}$ can be directly calculated from the surface S. Figure 2.4b shows the condition of the maximum ambiguity case.

RATIO MV TRANSFORMATION In [Ghasemi et al., 2012], Ghasemi et al. proposed a ratio membership value transformation method to calculate mass functions. The FCM algorithm was first used to generate MVs f_{ω_c} for each pixel. Then the MVs are used to build the mass functions. For this purpose, the three ratios of the available MVs are calculated, corresponding to three situations: "no uncertainty" (NU), "semi-uncertainty" (SU), and "perfect-uncertainty" (PU). First, PU is a critical situation in which the Ratio MVs are smaller than α, then the mass function is calculated as $m(\{\omega_1\}) = m(\{\omega_2\}) = m(\Omega) = (f_{\omega_1} + f_{\omega_2})/3$. Second, two thresholds α and β with $\alpha = 1.5$ and $\beta = 3$ are selected to control the boundary between SU and PU, and between NU and SU, separately. For example, with

$$f_{\omega_1} = 0.18, \quad f_{\omega_2} = 0.81, \quad RMV = f_{\omega_1}/f_{\omega_2} = 4.5, \quad RMV > \beta,$$

the two MVs fall in the NU category. If

$$f_{\omega_1} = 0.25, \quad f_{\omega_2} = 0.65, \quad RMV = f_{\omega_1}/f_{\omega_2} = 2.6, \quad \alpha < RMV < \beta,$$

Table 2.2: Example of credal partition.

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_5(A)$
$\{\emptyset\}$	1	0	0	0	0
$\{a\}$	0	0	1	0.5	0.6
$\{b\}$	0	0	0	0.3	0.4
$\{a, b\}$	0	1	0	0.2	0

the two MVs are in the SU category. The mass functions are calculated as

$$m(\{\omega_1\}) = f_{\omega_1} - \frac{\lambda_{\omega_1,\omega_2}}{2}, \tag{2.26a}$$

$$m(\{\omega_2\}) = f_{\omega_2} - \frac{\lambda_{\omega_1,\omega_2}}{2}, \tag{2.26b}$$

$$m(\Omega) = \lambda_{\omega_1,\omega_2}, \tag{2.26c}$$

where λ is an uncertainty distance value defined as $\lambda_{\omega_1,\omega_2} = \frac{|f_{\omega_1} - f_{\omega_2}|}{\beta - \alpha}$.

2.5.2.2 *Evidential C-means (ECM)*

In [Denœux et al., 2004], Denœux et al. proposed an evidential clustering algorithm, called EVCLUS, based on the notion of credal partition, which extends the existing concepts of hard, fuzzy (probabilistic), and possibilistic partition by allocating to each object a "mass of belief", not only to single clusters but also to any subsets of $\Omega = \{\omega_1, \dots, \omega_C\}$.

CREDAL PARTITION Assuming there is a collection of five objects for two classes, mass functions for each source are given in Table 2.2. They represent different situations: the mass function of object 1 indicates strong evidence that the class of object 1 does not lie in Ω; the class of object 2 is completely unknown, and the class of object 3 is known with certainty; the cases of objects 4 and 5 correspond to situations of partial knowledge (m_5 is Bayesian). The EVCLUS algorithm generates a credal partition for dissimilarity data by minimizing a cost function.

EVIDENTIAL C-MEANS (ECM) The ECM algorithm [Masson et al., 2008] is another method for generating a credal partition from data. In ECM, a cluster is represented by a prototype p_c. For each non-empty set $A_j \subseteq \Omega$, a prototype $\bar{p}_j$ is defined as the center of mass of the prototypes p_c such that $\omega_c \in A_j$. Then the non-empty focal set is defined as $F = \{A_1, \dots, A_f\} \subseteq 2^\Omega \setminus \{\emptyset\}$. Deriving a credal partition from object data implies determining, for each object x_i, the quantities $m_{ij} = m_i(A_j), A_i \neq \emptyset, A_j \subseteq \Omega$, in such a way that m_{ij} is low (resp. high) when the distance between x_i and the focal set $\bar{p}_j$ is high (resp. low). The distance between an object and any nonempty subset of Ω is then defined by

$$d_{ij}^2 = \left\|x_i - \bar{p}_j\right\|^2. \tag{2.27}$$

2.5.2.3 *Gaussian distribution (GD)-based model*

Besides the FCM-based and credal partition-based methods, the mass functions can also be generated from GD [Chen et al., 2012]. The mass of simple hypotheses $\{\omega_c\}$ can be obtained from the assumption of GD according to the information x_i of a pixel from an input image to cluster c as follows:

$$m(\{\omega_c\}) = \frac{1}{\sigma_c\sqrt{2\pi}} \exp \frac{-(x_i - \mu_c)^2}{2\sigma_c^2}, \tag{2.28}$$

where μ_c and σ_c^2 represent, respectively, the mean and the variance of the cluster c, which can be estimated by

$$\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_i, \tag{2.29}$$

$$\sigma_c^2 = \frac{1}{n_c} \sum_{i=1}^{n_c} (x_i - \mu_c)^2, \tag{2.30}$$

where n_c is the number of pixels in the cluster c. The mass of multiple hypotheses $\{\omega_1, \omega_2, \dots, \omega_T\}$ is determined as

$$m(\{\omega_1, \omega_2, \dots, \omega_T\}) = \frac{1}{\sigma_t\sqrt{2\pi}} \exp \frac{-(x_i - \mu_t)^2}{2\sigma_t^2}, \tag{2.31}$$

where $\mu_t = \frac{1}{T}\sum_{i=1}^{T} \mu_i$, $\sigma_t = \max(\sigma_1, \sigma_2, \dots, \sigma_T), 2 \leqslant T \leqslant C$, C is the number of clusters.

2.5.2.4 *Binary frames of Discernment (BFOD)*

Under the assumption that the membership value is available, Safranek et al. introduced a BFOD-based method [Safranek et al., 1990] to transform membership values into mass functions. The BFOD is constructed as $\Omega = \{\omega, \neg\omega\}$ with a function $cf(\nu)$, taking values in $[0, 1]$ that assigns confidence factors. According to the authors, the sigmoid and one-sided Gaussian functions are the most appropriate functions for defining $cf(\nu)$. Once a confidence value is obtained, the transformation into mass functions can be accomplished by defining appropriate transfer functions:

$$m(\{\omega\}) = \frac{B}{1-A} cf(\nu) - \frac{AB}{1-A}, \tag{2.32a}$$

$$m(\{\neg\omega\}) = \frac{-B}{1-A} cf(\nu) + B, \tag{2.32b}$$

$$m(\Omega) = 1 - m(\{\omega\}) - m(\{\neg\omega\}), \tag{2.32c}$$

where A and B are user-specific parameters. In (2.32a) and (2.32b), the left-hand side stays clamped at zero when the right-hand side goes negative. Parameter A is the confidence-factor axis intercept of the curve that depicts the dependence of $m(\{\omega\})$ on confidence factors, and B is the maximum support value assigned to $m(\{\omega\})$ or $m(\{\neg\omega\})$.

2.6 conclusion

BFT is a powerful tool to represent imperfect information and combine different sources of information by Dempster's rule. In this chapter, we first recalled how imperfect information can be represented by mass functions and how multiple sources of evidence can be combined by Dempster's rule. We then introduced the discounting operation for unreliable evidence and decision-making based on mass functions. In addition, we gave a summary of existing basic belief assignment methods that have been used in medical image segmentation. In the next chapter, we will provide an overall review of BFT-based medical image segmentation methods.

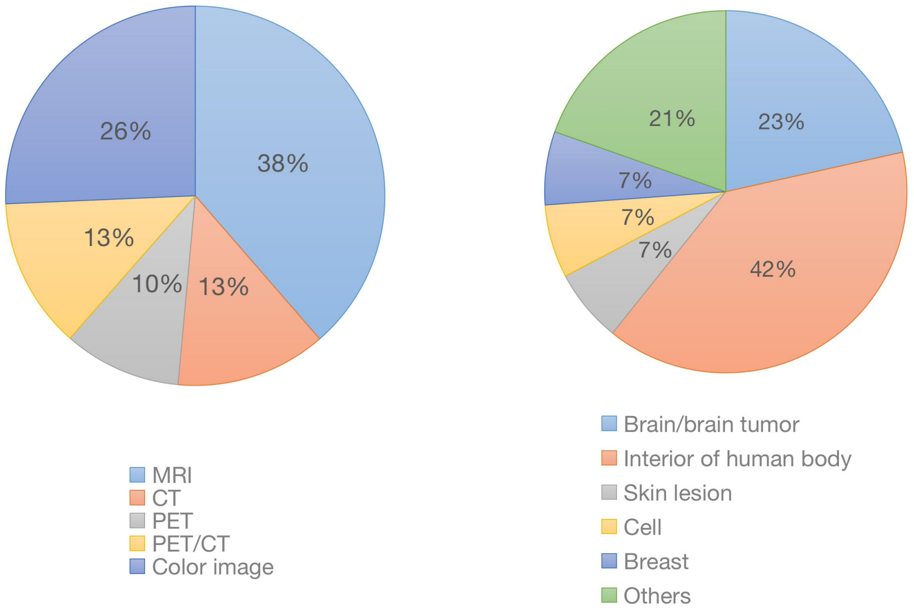
Chapter 3

Review of BFT-based medical image segmentation methods

Contents

To summarize the BFT-based medical image segmentation methods, we can either classify them by the input modality of the images or by the specific clinical application. Figure 3.1a shows the proportion of types of medical images applied in the segmentation task and Figure 3.1b displays the proportions of application in the medical image segmentation task.

We have presented the common basic belief assignment methods to generate mass functions for medical images in Section 2.5. The most interesting way now to analyze those segmentation methods is to classify them according to the stage at which they fuse mass functions. Figure 3.2 gives an overview of the BFT-based medical image segmentation methods classified according to the fusion operation they perform. We first classify methods by the number of classifiers/clusterers, i.e., whether the fusion of mass functions is performed at the classifiers/clusterers level. Then we consider the number of input modalities, i.e., whether the fusion of mass functions is performed at the modality level. It should be noted that the medical image segmentation with single-modality input and single classifier/clusterer can also have a fusion operation, which is performed at the pixel/voxel level. Figure 3.3 gives the proportions of the BFT-based segmentation methods if multiple classifiers/clusterers or multimodal medical images are used. 73% of the methods use a single classifier or cluster. Among these methods, 24% take single-modality medical images as input,



(a) **(b)**

Figure 3.1: (a) The composition of input modality of the medical image, (b)The composition of specific medical application.

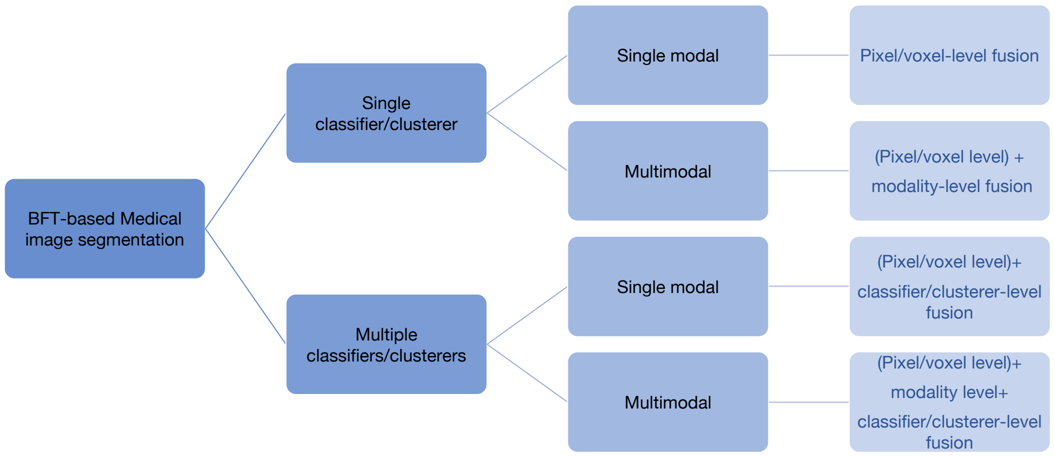

BFT-based Medical image segmentation
Single classifier/clusterer
Multiple classifiers/clusterers
Single modal
Multimodal
Single modal
Multimodal
Pixel/voxel-level fusion
(Pixel/voxel level) + modality-level fusion
(Pixel/voxel level)+ classifier/clusterer-level fusion
(Pixel/voxel level)+ modality level+ classifier/clusterer-level fusion

Figure 3.2: Overview of BFT-based medical image segmentation methods.

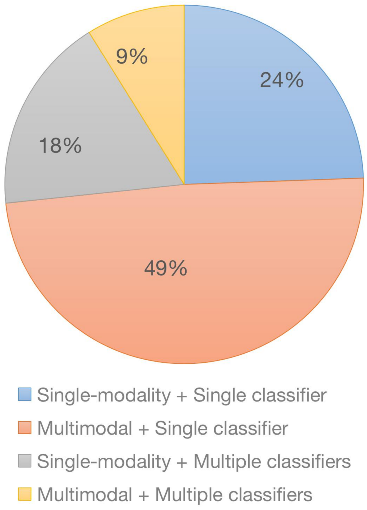


Figure 3.3: The composition of the proportion of modalities and classifiers/clusterers.

and 49% take multimodal medical images as input. The remaining 27% of the methods use several classifiers or clusterers. Among those methods, 18% take single-modality medical images as input, against 9% for multimodal medical images. In Sections 3.1 and 3.2 we give more details about the fusion operations and discuss their performances.

3.1 SEGMENTATION WITH A SINGLE CLASSIFIER OR CLUSTERER

The medical image segmentation methods summarized in this section focus on using a single classifier or clusterer. To simplify the introduction, we classify them into single-modality and multimodal inputs according to whether the authors treat the input images as single or multiple source inputs. We will introduce them in Sections 3.1.1 and 3.1.2, respectively.

3.1.1 *Single-modality evidence fusion*

Figure 3.4 shows the framework of single-modality evidence fusion with a single classifier (we only take the classifier as an example in the rest of this chapter to simplify the explanation). The inputs of the framework are single-modality medical images. The segmentation process comprises three steps: mass calculation (including feature extraction and BBA), feature-level mass fusion, and decision-making. Since decision-making is not the focus of this chapter, we will not go into details about it; readers can refer to Ref. [Denœux, 2019a] for a recent review of decision methods based on belief functions.

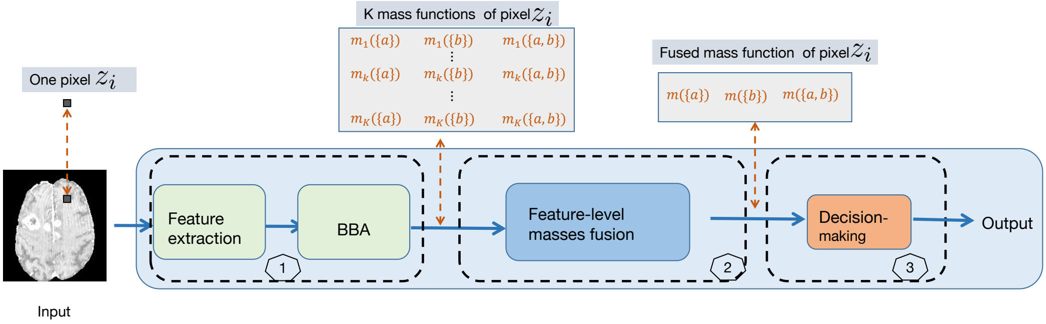

K mass functions of pixel z_i
$m_1(\{a\})$ $m_1(\{b\})$ $m_1(\{a,b\})$
$m_k(\{a\})$ $m_k(\{b\})$ $m_k(\{a,b\})$
$m_K(\{a\})$ $m_K(\{b\})$ $m_K(\{a,b\})$
One pixel z_i
Fused mass function of pixel z_i
$m(\{a\})$ $m(\{b\})$ $m(\{a,b\})$
Feature extraction
BBA
Feature-level masses fusion
Decision-making
Output
Input
1
2
3

Figure 3.4: Example framework of single-modality evidence fusion with single classifier. The segmentation process is composed of three steps: (1) image information is transferred into the feature extraction block, and some BBA methods are used to get feature-level mass functions; (2) Dempster's rule is used to fuse feature-level mass functions; (3) decision-making is made based on the fused mass functions to output a segmented mask. We take only one pixel z_i as an example and show how BFT works on pixel-level evidence fusion under a binary segmentation task to simplify the process. For each pixel z_i, we can obtain K mass functions from BBA. After feature-level evidence fusion, a fused mass function is assigned to pixel z_i to represent the degree of belief this pixel belongs to classes a, b and ignorance.

(1) First, the mass calculation step assigns each pixel/voxel K mass functions based on the evidence of K different BBA methods, K input features, K nearest neighbors, or K prototype centers.

(2) Dempster's rule is then used for each pixel/voxel to fuse the feature-level mass functions.

(3) Finally, decisions are made to obtain final segmentation results.

Here, a feature extraction method is used to generate MVs (corresponding to traditional medical image segmentation methods) or deep features (corresponding to deep learning-based medical image segmentation methods). The feature extraction method could be intensity-based methods such as threshold intensity, probabilistic-based methods such as SVM, fuzzy set-based methods such as FCM, etc. There are various basic belief assignment methods, therefore we introduce them in specific tasks in the following. In general, the methods introduced in this section focus on feature-level evidence fusion. Table 3.1 shows the segmentation methods with a single-modality input and classifier/cluster that focus on feature-level evidence fusion.

In [Suh et al., 1993], Suh et al. developed a knowledge-based endocardial and epicardial boundary detection and segmentation system with cardiac MR image sequences. Pixel and location information were mapped into mass functions by Shafer's model [Shafer, 1976] (see Section 2.5.1.1).

Table 3.1: Summary of BFT-based medical image segmentation methods with single-modality inputs and a single classifier/clusterer.

Publications	Input image type	Application	BBA method
[Suh et al., 1993]	MR images	cardiac segmentation	Shafer's model
[Gerig et al., 2000]	MR images	brain tissue segmentation	BFOD
[Vauclin et al., 2005]	CT	lung and spinal canal segmentation	Shafer's model
[Vannoorenberghe et al., 2006]	CT	thoracic segmentation	EKNN
[Derraz et al., 2013b]	optical imaging with color	cell segmentation	Appriou's model
[Derraz et al., 2013a]	optical imaging with color	retinopathy segmentation	Appriou's model
[Derraz et al., 2013b]	optical imaging with color	retinopathy segmentation	Appriou's model
[Lian et al., 2017b]	FDG-PET	lung tumor segmentation	ECM
[Lian et al., 2017c]	FDG-PET	lung tumor segmentation	ECM
[Huang et al., 2021c]	PET-CT	lymphoma segmentation	ENN
[Huang et al., 2022b]	PET-CT	lymphoma segmentation	ENN, RBF

The mass functions from the two sources were fused by using Dempster's rule (see Section 2.2). Experiments were applied to cardiac short-axis images and obtained an excellent success rate ($> 90\%$). However, this work only focused on cardiac boundary detection and did not discuss the details of the heart. In [Vauclin et al., 2005], Vauclin et al. proposed a BFT-based lung and spinal canal segmentation model. The k-means clustering algorithm was first used to perform a pre-segmentation. Then a 3D filter exploits the results of the pre-segmentation to compute the MVs from spatial neighbors using Shafer's model. Segmentation results showed the credal partition permits the reduction of the connection risks between the trachea and the lung when they are very close and between the left and right lungs at the anterior or posterior junctions.

In [Gerig et al., 2000], Gerig et al. presented a method for automatic segmentation and characterization of object changes in time series of 3D MR images. A set of MVs was derived from time series according to brightness changes. The BFOD transformation (see Section 2.5.2.4) was used to map the obtained features into mass functions. Then the set of evidence was combined by Dempster's rule (See Section 2.2). A comparison with results from alternative segmentation methods revealed an excellent sensitivity and specificity performance in the brain lesion region. The author also pointed out that better performance could be obtained with multimodal and multiple time-series evidence fusion. Simulation results showed that about 80% of the implanted voxels could be detected for most generated lesions.

In [Vannoorenberghe et al., 2006], Vannoorenberghe et al. presented a BFT-based thoracic image segmentation method. First, a K-means algorithm performed coarse segmentation on the original CT images. Second, the EKNN rule (see Section 2.5.1.2) was applied by considering spatial information and calculating feature-level mass functions. The authors claimed that using this segmentation scheme leads to a complementary approach combining region segmentation and edge detection. Experimental results showed promising results on 2D and 3D CT images for lung segmentation.

In [Derraz et al., 2013b], Derraz et al. proposed an active contour (AC)-based [Chan et al., 2000] global segmentation method for vector-valued image that incorporated both probability and mass functions. All features issued from the vector-valued image were integrated with inside/outside descriptors to drive the segmentation results by maximizing the Maximum-Likelihood distance between foreground and background. Appriou's second model (2.15) was used to calculate the imprecision caused by low contrast and noise between inside and outside descriptors issued from the multiple channels. Then the fast algorithm based on Split Bregman [Goldstein et al., 2009] was used for final segmentation by forming a fast and accurate minimization algorithm for the Total Variation (TV) problem. Experiments were conducted on color biomedical images (eosinophil, lymphocyte, eosinophil, monocyte, and neutrophil cell [Mohamed et al., 2012]) and achieve around 6% improvements by using F-score on five image groups. In the same year, Derraz et al. proposed a new segmentation method [Derraz et al., 2013a] based on Active Contours (AC) for the vector-valued image that incorporates Bhattacharyya's distance [Michailovich et al., 2007]. The only difference is that the authors calculate the probability functions by Bhattacharyya distance instead of the Maximum-Likelihood distance in this paper. The performance of the proposed algorithm was demonstrated on the retinopathy dataset [Quellec et al., 2008; Niemeijer et al., 2009], showing an increase of 3% in F-score compared with the best-performed methods.

In [Lian et al., 2017b], Lian et al. introduced a tumor delineation method in fluorodeoxyglucose positron emission tomography (FDG-PET) images by using spatial ECM [Lelandais et al., 2014] with adaptive distance metric, a variant of the ECM algorithm recalled in Section 2.5.2.2. The authors proposed the adaptive distance metric to extract the most valuable features, and spatial ECM was used to calculate mass functions. Compared with ECM and spatial ECM, the proposed method showed a 14% and 10% increase in Dice score when evaluated on the FDG-PET images of non-small cell lung cancer (NSCLC) patients, which constitutes a very good performance.

In [Huang et al., 2021c], we proposed a 3D PET/CT lymphoma segmentation framework with BFT and deep learning. In this paper, the PET and CT images were concatenated as a signal modal input method and transferred into UNet to get high-level semantic features. Then the ENN classifier (see Section 2.5.1.2) was used to map high-level semantic features into mass functions by fusing the contribution of K prototypes. Moreover, the segmentation uncertainty was considered in this paper with an uncertainty loss during training. Based on the first work, we verified the similarity of radial basis function (RBF, see Section 5.2.1.2) network and ENN in uncertainty quantification and merged them with the deep neural network (UNet, see Section 1.3.1) for lymphoma segmentation [Huang et al., 2022b]. The segmentation performance confirmed that RBF is an alternative approach of ENN to act as an evidential classifier and showed that the proposal outperforms the baseline methods and the state-of-the-art both in accuracy and reliability. We will introduce more detail about this work in Chapter 5.

Before 2020, the BFT-based medical image segmentation methods with single-modality medical used low-level image features, e.g., grayscale and shape features, to generate mass functions, which limits the segmentation accuracy. However, none of them discussed the reliability of the segmentation results. We are the first to merge BFT with a deep segmentation model (UNet) and optimize the mass functions with some learning algorithms [Huang et al., 2021c]. Based on this, we further discuss the relationship between segmentation accuracy and reliability in [Huang et al., 2022b], which takes a new direction to study reliable medical image segmentation methods and bridge the gap between experimental results and clinical application.

3.1.2 *Multimodal evidence fusion*

Single-modality medical images often do not contain enough information to present the information about the disease and are often tainted with uncertainty. In addition to feature-level evidence fusion, the fusion of multimodal evidence is also important to achieve accurate and reliable medical image segmentation performance. Approaches for modality-level evidence fusion can be summarized into three main categories according to the way they calculate the evidence: probabilistic-based fusion, fuzzy set-based fusion, and BFT-based fusion. The development of convolution neural networks (CNNs) further contributes to the probabilistic-based fusion methods [Zhou et al., 2019], which can be summarized into image-level fusion (e.g., data concatenation [Peiris et al., 2021]), feature-level fusion (e.g., attention mechanism concatenation [Zhou et al., 2020; Zhou et al., 2022]), and decision-level fusion (e.g., model ensembles [Kamnitsas et al., 2017]). However, none of those methods considers the conflict between sources of evidence, i.e., the modality-level uncertainty is not well studied, which limits the reliability and explainability of the performance.

This section focuses on the BFT-based segmentation methods with modality-level evidence fusion. Figure 3.5 shows an example framework of multimodal evidence fusion with a single classifier. We separate the segmentation process into four steps: mass calculation, feature-level evidence fusion (optional), modality-level evidence fusion, and decision-making. Compared with single-modality evidence fusion reviewed in Section 3.1.1, multimodal evidence fusion focuses here not only on feature-level but also on modality-level evidence fusion. It should be noted that feature-level evidence fusion is not necessary in this case. Multimodal evidence fusion

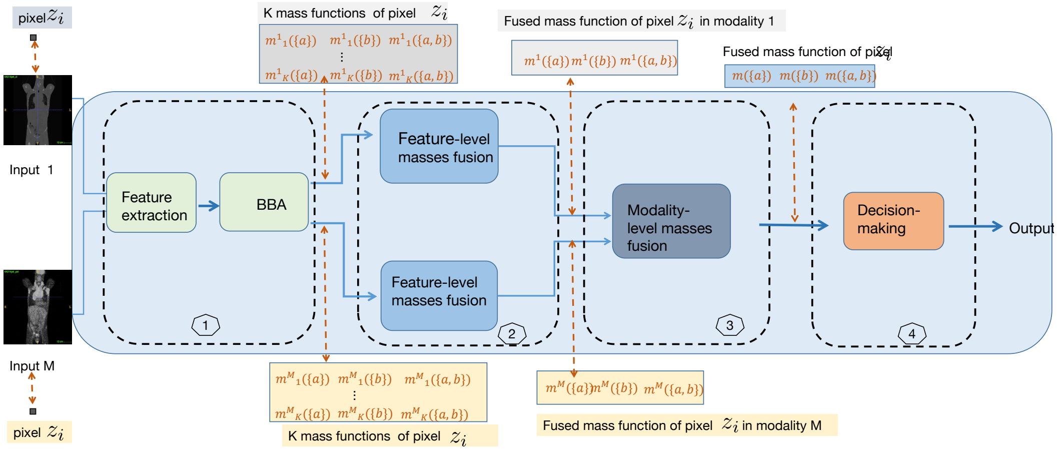

pixel z_i
K mass functions of pixel z_i
$m^1{}_1(\{a\})$ $m^1{}_1(\{b\})$ $m^1{}_1(\{a,b\})$
⋮
$m^1{}_K(\{a\})$ $m^1{}_K(\{b\})$ $m^1{}_K(\{a,b\})$
Fused mass function of pixel z_i in modality 1
$m^1(\{a\})m^1(\{b\})$ $m^1(\{a,b\})$
Fused mass function of pixel z_i
$m(\{a\})$ $m(\{b\})$ $m(\{a,b\})$
Input 1
Feature extraction
BBA
Feature-level masses fusion
Feature-level masses fusion
Modality-level masses fusion
Decision-making
Output
1
2
3
4
Input M
pixel z_i
$m^M{}_1(\{a\})$ $m^M{}_1(\{b\})$ $m^M{}_1(\{a,b\})$
⋮
$m^M{}_K(\{a\})$ $m^M{}_K(\{b\})$ $m^M{}_K(\{a,b\})$
K mass functions of pixel z_i
$m^M(\{a\})m^M(\{b\})$ $m^M(\{a,b\})$
Fused mass function of pixel z_i in modality M

Figure 3.5: Example framework of multimodal evidence fusion with a single classifier. The segmentation process is composed of four steps: (1) for each modal input, image features are fed into the classifier, and some BBA methods are used to get feature-level mass functions; (2) inside each modal, the calculated feature-level mass functions are fused by using Dempster's rule to generate modality-level mass functions; (3) between the modalities, the calculated modality-level mass functions are fused by using Dempster's rule again; (4) decision-making is made based on the fused mass functions to output a segmented mask.

is the most popular application for BFT in the medical image segmentation domain. Therefore we classify those methods according to their input modal for better analysis. Table 3.2 summarizes the segmentation methods with multimodal inputs and a single classifier/cluster with the main focus on modality-level evidence fusion.

FUSION OF MULTIMODAL MR IMAGES In [Bloch, 1996], Bloch first proposed a BFT-based dual-echo MR pathological brains tissue segmentation model with uncertainty and imprecision quantification. The author assigned mass functions based on a reasoning approach that uses gray-level histograms provided by each image to choose focal elements. After defining mass functions, Dempster's rule combined the mass from dual-echo MR images for each pixel. Based on the first work with BFT, in [Bloch, 2008], Bloch proposed to use fuzzy mathematical morphology [Bloch, 1995], i.e., erosion and dilation, to generate mass functions by introducing imprecision in the probability functions and estimating compound hypotheses. Then Dempster's rule (see Section 2.2) is used to fuse mass functions from multimodal images. It should be noted that in this paper, the strong assumption is made that it is possible to represent imprecision by a fuzzy set, also called the structuring element. Application examples on dual-echo MR image fusion showed that the fuzzy mathematical morphology operations could represent either spatial domain imprecision or feature space imprecision (i.e., gray levels features).

Table 3.2: Summary of BFT-based medical image segmentation methods with multimodal input and single classifier/clusterer.

Publications	Input image type	Application	BBA method
[Vannoorenberghe et al., 1999]	optical imaging with color	skin cancer segmentation	GD-based model
[Bloch, 1996]	multimodal MR images	brain tissue segmentation	prior knowledge
[Taleb-Ahmed et al., 2002]	multimodal MR images	vertebrae segmenta	threshold+contour distance
[Chaabane et al., 2009]	optical imaging with color	cell lesion segmentation	Appriou's model
[Zhu et al., 2002]	multimodal MR images	brain tissue segmentation	Zhu's model
[Bloch, 2008]	multimodal MR images	brain tissue segmentation	prior knowledge
[Chaabane et al., 2011]	optical imaging with color	cell lesion segmentation	Aprrious's method
[Ghasemi et al., 2012]	multimodal MR images	brain tissue segmentation	Ratio MV
[Harrabi et al., 2012]	optical imaging with color	breast cancer segmentation	GD-based model
[Lelandais et al., 2012]	PET	biological target tumor segmentation	ECM
[Wang et al., 2013]	multimodal MR images	cerebral infraction segmentation	Zhu's model
[Ghasemi et al., 2013]	multimodal MR images	brain tumor segmentation	Ratio MV
[Lelandais et al., 2014]	multi-tracer PET	biological target tumor segmentation	ECM
[Makni et al., 2014]	multi-parametric MR image	prostate zonal anatomy	ECM
[Derraz et al., 2015]	PET/CT	lung tumor segmentation	Appriou's model
[Trabelsi et al., 2015]	optical imaging with color	skin lesion segmentation	None
[Xiao et al., 2017]	multimodal MR images	vascular segmentation	GD-based model
[Lian et al., 2017a]	FDG-PET/CT	lung tumor segmentation	ECM
[Lian et al., 2018b]	FDG-PET/CT	lung cancer segmentation	ECM
[Tavakoli et al., 2018]	multimodal MR images	brain tissue segmentation	Ratio MV
[Lima et al., 2019]	multimodal MR images	brain tissue segmentation	Ratio MV
[Lian et al., 2018a]	PET/CT	lung tumor segmentation	ECM

The visualized brain tissue segmentation results show the robustness of the proposed method.

As we mentioned in Section 2.5, Zhu et al. proposed modeling mass functions in BFT using FCM and spatial neighborhood information for image segmentation [Zhu et al., 2002]. The visualized segmentation results on MR brain images showed that the fusion-based segmented regions are relatively homogeneous, enabling accurate measurement of brain tissue volumes compared with single-modality input MR image input.

In [Ghasemi et al., 2012], Ghasemi et al. presented a brain tissue segmentation approach based on FCM and BFT. The authors used the FCM to model two different input features: pixel intensity and spatial information, as membership values (MVs). Then for each pixel, the Ratio MV transformation (see Section 2.5.2.1) was used to map MVs into mass functions. Last, the authors used Dempster's rule (see Section 2.2) to fuse intensity-based and spatial-based mass functions to get final segmentation results. Compared with FCM, the authors reported an increase in Dice score and accuracy. As an extension of [Ghasemi et al., 2012], Ghasemi et al. proposed an unsupervised brain tumor segmentation method that modeled pixel intensity and spatial information into mass functions with Ratio MV transformation and fused the two mass functions with Dempster's rule in [Ghasemi et al., 2013].

In [Wang et al., 2013], Wang et al. proposed a lesion segmentation method for infarction and cytotoxic brain edema. The authors used a method similar to Zhu's model (see Section 2.5.2.1) to define simple and double hypotheses. FCM was used first to construct the mass functions of simple hypotheses $\{a\}$ and $\{b\}$. Then masses were assigned to double hypotheses as $m(\{a,b\}) = \frac{1}{4} \times \frac{\min(m(\{a\}),m(\{b\}))}{m(\{a\})+m(\{b\})}$. Finally, the authors used Dempster's rule for modality-level evidence fusion. The results showed that infarction and cytotoxic brain edema around the infarction could be well segmented by final segmentation.

In [Makni et al., 2014], Makni et al. introduced a multi-parametric MR image segmentation model by using spatial neighborhood in ECM for prostate zonal anatomy. The authors extended the ECM (see Section 2.5.2.2) with neighboring voxels information to map multi-parametric MR images into mass functions. Then they used prior knowledge related to defects in the acquisition system to reduce uncertainty and imprecision. Finally, the authors used Dempster's rule to fuse the mass functions from the multi-parametric MR images. The method achieved good performance on prostate multi-parametric MR image segmentation.

In [Tavakoli et al., 2018], Tavakoli et al. proposed a segmentation method based on the evidential fusion of different modalities (T1, T2, and Proton density (PD)) for brain tissue. The authors used FCM to get MVs and used the Ratio MV transformation (see Section 2.5.2.1) to transform the clustering MVs into mass functions. The authors first formed the belief structure for each modal image and used Dempster's rule to fuse the three modalities' mass functions of T1, T2, and PD. Compared with FCM, this method achieved a 5% improvement in the Dice score. Based on Tavakoli's method [Tavakoli et al., 2018], Lima et al. proposed a modified brain tissue segmentation method in [Lima et al., 2019]. The authors tested their method with four modality inputs: T1, T2, PD, and Flair. The reported results outperformed both the baseline method FCM and Tavakoli's method with four-modality evidence fusion.

In [Xiao et al., 2017], Xiao et al. proposed an automatic vascular segmentation algorithm, which combines the grayscale and shape features of blood vessels and extracts 3D vascular structures from the head phase-contrast MR angiography dataset. First, grayscale and shape features are mapped into mass functions by using the GD-based method (see Section 2.5.2.3). Second, a new reconstructed vascular image was established according to the fusion of vascular grayscale and shape features by Dempster's rule. Third, a segmentation ratio coefficient was proposed to control the segmentation result according to the noise distribution characteristic. Experiment results showed that vascular structures could be detected when both grayscale and shape features are robust. Compared with traditional grayscale feature-based or shape feature-based methods, the proposal showed better performance in segmentation accuracy with the decreased over-segmentation and under-segmentation ratios by fusing two sources of information.

Since Bloch's early work fully demonstrated the advantages of BFT in modeling uncertain and imprecision, introducing partial or global ignorance, and fusing conflicting evidence in a multimodal MR images segmentation task, researchers in this domain have gone further to explore the advantages of BFT in multi MR image fusion. Among these research works, FCM is the most popular clustering algorithm to map input information into MVs. Ration MV transformation or Zhu's model (see Section 2.5.2.1) is usually used to generate mass functions from MVs. Apart from these two-step methods, the GD-based model (see Section 2.5.2.3) can also be used to generate mass functions directly.

FUSION OF RGB CHANNELS In [Vannoorenberghe et al., 1999], Vannoorenberghe et al. pointed out that taking the R, G, and B channels as three independent information sources can be limited and nonoptimal for medical image tasks and proposed a color image segmentation method based on BFT. They calculated the degree of evidence by mapping R, G, and B channel intensity into mass functions using the Gaussian distribution information (similar to GD-based model) with an additional discounting operation). Then three pieces of evidence were fused with Dempster's rule (see Section 2.2). The proposed segmentation method was applied to biomedical images to detect skin cancer (melanoma). Experiments showed a significant part of the lesion is correctly extracted from the safe skin. The segmentation performance is limited by feature representation, e.g., some regions correspond to pixels that cannot be classified as either the safe skin or the lesion because only the pixel-level feature is insufficient for hard-example segmentation.

In [Chaabane et al., 2009] Chaabane et al. proposed a color medical image segmentation method based on fusion operation. Compared to [Vannoorenberghe et al., 1999], the authors first modeled probabilities for each region by a Gaussian distribution, then used Appriou's second model (2.15) to map probability into mass functions. Dempster's rule then combined the evidence from the three color channels. Compared with single-channel segmentation results, the fused results achieved a 10% increase in segmentation sensitivity.

Different from the methods described in [Vannoorenberghe et al., 1999] and [Chaabane et al., 2009] that decompose color images into R, G, B channels, Harrabi et al. [Harrabi et al., 2012] presented a color image segmentation method that represents the color image with 6 color spaces (RGB, HSV, YIQ, XYZ, LAB, LUV). The segmentation operation is based on multi-level thresholding and evidence fusion techniques. First, the authors identified the most significant peaks of the histogram by multi-level thresholding with the two-stage Otsu optimization approach. Second, the GD-based model (see Section 2.5.2.3) was used to calculate the mass functions for each color space. Then the authors used Dempster's rule to combine six sources of information. Compared with single color spaces, such as RGB and HSV, the combined result taking into account six color spaces, has a significant increase in segmentation sensitivity, for example, an increase of 4% and 7% as compared to RGB and HSV, respectively.

In [Trabelsi et al., 2015], Trabelsi et al. applied BFT in optical imaging with color to improve skin lesion segmentation performance. The authors decomposed color images into R, G, and B channels and applied the FCM method on each channel to get probability functions for pixel x in each color space. In this paper, the authors take the probability functions as mass functions and calculate the orthogonal sum of the probability functions from the three-channel images. Even though experiments showed about 10% improvements in segmentation accuracy compared with single-channel results, this work does not harness the full power of BFT as it only considers Bayesian mass functions.

In general, the BFT-based RGB medical image segmentation approaches are used to generate mass functions from possibility distributions, e.g., Gaussian distribution and Possibility C-means distribution, and fuse them by Dempster's rule. Though the authors claimed they could get better performance compared with single color input, the segmentation performance is limited by features because gray-scale and intensity features are not robust and efficient in representing image information. Further research could take both feature extraction and uncertainty quantification into consideration, e.g., a deep feature extraction model with an evidential classifier, to improve the performance.

FUSION OF PET/CT In [Lelandais et al., 2012], Lelandais et al. proposed a BFT-based multi-trace PET images segmentation method to segment biological target volumes. First, the authors used a modified FCM algorithm with the discounting algorithm to determine mass functions. The modification integrated a disjunctive combination of neighboring voxels inside the iterative process. Second, the operation of reduction of imperfect data was conducted by fusing neighboring voxels using Dempster's rule (see Section 2.2). Based on this first work, Lelandais et al. proposed an ECM-based fusion model [Lelandais et al., 2014] for biological target volume segmentation with multi-tracer PET images. The segmentation method introduced in this paper is similar to the one introduced in [Makni et al., 2014] with a different application.

In [Derraz et al., 2015], Derraz et al. proposed a multimodal tumor segmentation framework for PET and CT inputs. Different from Lelandais et al.'s work that uses ECM or optimized ECM to generate mass functions directly, the authors construct mass functions in two steps. They first proposed a NonLocal Active Contours (NLAC) based variational segmentation framework to get probability results. Then, similar to the authors' previous work [Derraz et al., 2013b; Derraz et al., 2013a], they used Appriou's second model (2.15) to map MVs into mass functions. Then Dempster's rule was used to fuse the mass functions from PET and CT modalities. The framework was evaluated on a lung tumor segmentation task. Compared with the state-of-the-art methods, this framework yielded the highest Dice score for tumor segmentation.

Based on Lelandais et al.'s work, Lian et al. proposed a tumor segmentation method [Lian et al., 2017a] using Spatial ECM [Lelandais et al., 2014] with Adaptive Distance Metric [Lian et al., 2017b] in FDG-PET images with the guidance of CT images. Based on the first work, Lian et al. proposed a co-segmentation method of lung tumor segmentation [Lian et al., 2018a; Lian et al., 2018b]. They took PET and CT as independent inputs and used ECM (see Section 2.5.2.2) to generate mass functions. At the same time, an adaptive distance metric was used to quantify clustering distortions and spatial relationships during the evidential clustering procedure. Then Dempster's rule was used to fuse mass functions from PET and CT modalities. The quantitative and qualitative evaluation results showed superior performance compared with single-modality segmentation results with an increase of 1% and 58% in PET and CT in Dice scores, respectively.

ECM is the most common BBA method to generate mass functions for PET/CT medical image segmentation approaches. Similar to the BFT-based RGB medical image segmentation methods, the segmentation performance here is limited by feature extraction methods. Further research could build on recent advancements in deep feature representation and combine ECM with deep neural networks to learn mass feature representation.

3.2 SEGMENTATION WITH SEVERAL CLASSIFIERS OR CLUSTERERS

It is common practice that two or more physicians cooperate for disease diagnosis, which can minimize the impact of physicians' misjudgments. Similarly, combining the results from multiple decision mechanisms as well as addressing the conflicts is critical to achieving a more reliable diagnosis. This section introduces the BFT-based medical image segmentation methods with several classifiers or clusterers. We follow the same approach as in Section 3.1 and separate the methods into single-modality and multimodal evidence fusion reviewed, respectively, in Sections 3.2.1 and 3.2.2.

3.2.1 *Single-modality evidence fusion*

Compared with the methods presented in Section 3.1.1, the methods introduced in this section focus on feature-level and classifier-level evidence fusion, which aims to minimize the impact of misjudgments caused by a single model's inner shortcomings. Figure 3.6 shows an example of a medical image segmentation framework with single-modality input and several classifiers. We separate the segmentation process into four steps: mass calculation, feature-level evidence fusion (optional), classifier-level evidence fusion, and decision-making. Similar to Figure 3.4, we assume for each

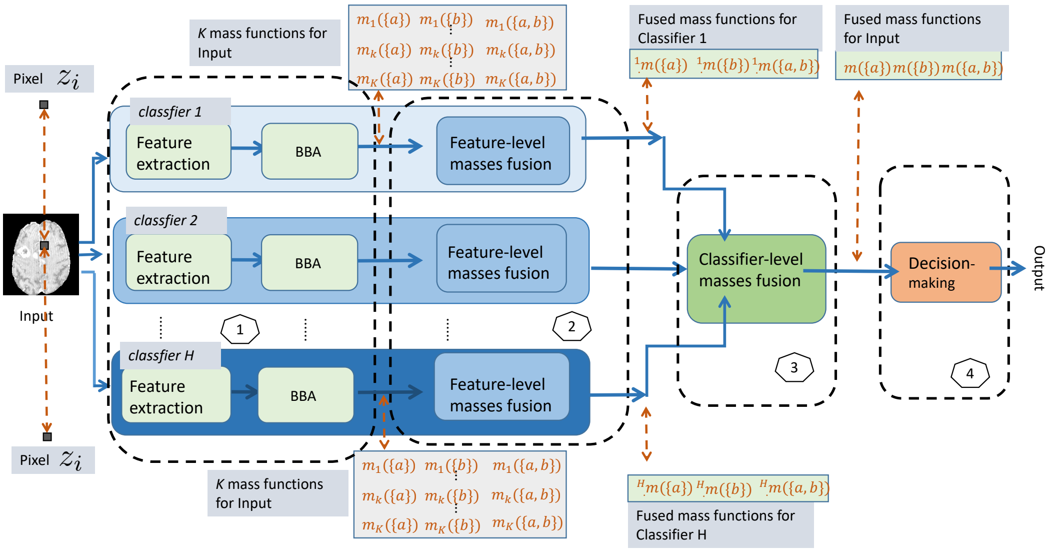

K mass functions for Input
$m_1(\{a\})$ $m_1(\{b\})$ $m_1(\{a,b\})$
$m_k(\{a\})$ $m_k(\{b\})$ $m_k(\{a,b\})$
$m_K(\{a\})$ $m_K(\{b\})$ $m_K(\{a,b\})$
Fused mass functions for Classifier 1
${}^1m(\{a\})$ ${}^1m(\{b\})$ ${}^1m(\{a,b\})$
Fused mass functions for Input
$m(\{a\})$ $m(\{b\})$ $m(\{a,b\})$
Pixel z_i
classfier 1
Feature extraction
BBA
Feature-level masses fusion
classfier 2
Feature extraction
BBA
Feature-level masses fusion
Input
1
2
classfier H
Feature extraction
BBA
Feature-level masses fusion
Classifier-level masses fusion
3
Decision-making
4
Output
Pixel z_i
K mass functions for Input
$m_1(\{a\})$ $m_1(\{b\})$ $m_1(\{a,b\})$
$m_k(\{a\})$ $m_k(\{b\})$ $m_k(\{a,b\})$
$m_K(\{a\})$ $m_K(\{b\})$ $m_K(\{a,b\})$
${}^Hm(\{a\})$ ${}^Hm(\{b\})$ ${}^Hm(\{a,b\})$
Fused mass functions for Classifier H

Figure 3.6: Example framework of single-modality evidence fusion with several classifiers. The segmentation process is composed of four steps: (1) for each pixel, image features are transferred into the different classifiers, and some BBA methods are used to get feature-level mass functions; (2) inside each classifier, the calculated feature-level mass functions are fused by using Dempster's rule to generate classifier-level mass functions; (3) between the classifiers, the calculated classifier-level mass functions are fused by using Dempster's rule again; (4) decision-making is made based on the final fused mass functions and output a segmented mask.

Table 3.3: Summary of BFT-based medical image segmentation methods with multimodal inputs and several classifiers/clusterers.

Publications	Input image type	Application	BBA method
[Capelle et al., 2002]	MR images	brain tumor segmentation	EKNN
[Capelle et al., 2004]	MR images	brain tumor segmentation	EKNN+Shafer's model+Appriou's model
[Barhoumi et al., 2007]	optical imaging with color	skin lesion malignancy tracking	None
[Guan et al., 2011]	MR images	brain tissue segmentation	Zhu's model
[Ketout et al., 2012]	optical imaging with color	endocardial contour detection	threshold
[Wen et al., 2018]	MR images	brain tissue segmentation	Zhu's model+ GD-based model
[George et al., 2020]	optical imaging with color	breast cancer segmentation	Discounting
[Huang et al., 2021a]	PET-CT	lymphoma segmentation	ENN

pixel, we can obtain K mass functions by one classifier. After feature-level evidence fusion, for each pixel z_i, we can get H mass functions corresponding to H classifiers. Then we fuse the H mass functions and assign a fused mass function to the pixel z_i, representing the degree of belief this pixel belongs to classes a, b and ignorance. Table 3.3 summarizes the segmentation methods with single-modality inputs and several single classifiers/clusterers with the main focus on classifier/clusterer level evidence fusion.

In [Capelle et al., 2002], Capelle et al. proposed a segmentation scheme for MR images based on BFT. The authors used the Evidential K-NN rule recalled in Section 2.5.1.2 to map image features into mass functions. Then, they applied the evidential fusion process to classify the voxels. Based on this first work, Capelle et al. later proposed an evidential segmentation scheme of multimodal MR images for brain tumor detection in [Capelle et al., 2004]. This work focused on analyzing different evidential modeling techniques and on the influence of the introduction of spatial information to find the most effective brain tumor segmentation method. Three different BFT-based models: the distance-based BFT model (EKNN, see Section 2.5.1.2), the likelihood function-based BFT method (Shafer's model, see Section 2.5.1.1), and Appriou's first model (2.14c) were used to model information; Dempster's rule (see Section 2.2) was then used to combine the three mass functions. This study concluded that neighborhood information increases the evidence of class membership for each voxel, thus making the final decision more reliable. Experimental results showed better segmentation performance compared with the state-of-the-art methods when the paper was published.

In [Taleb-Ahmed et al., 2002], Taleb-Ahmed proposed a segmentation method for MR sequences of vertebrae in the form of images of their multiple slices with BFT. The authors used three different classifiers to calculate three kinds of mass functions. Firstly, the authors used gray-level intensity and standard deviation information to calculate two pixel-level mass functions with two fixed thresholds. Then the distance between two matching contours of consecutive slices (P and Q) was used to calculate contour-level mass functions as follows:

$$m(\{S_{PQ}\}) = \begin{cases} 1 - e^{-\eta|d(P_i,Q_i)-\beta|} & \text{if } d(P_i,Q_i) \in [\rho,\beta], \\ 0 & \text{otherwise} \end{cases} \tag{3.1a}$$

$$m(\overline{\{S_{PQ}\}}) = \begin{cases} 1 - e^{-\eta|d(P_i,Q_i)-\beta|} & \text{if } d(P_i,Q_i) \in (\beta,+\infty), \\ 0 & \text{otherwise} \end{cases} \tag{3.1b}$$

$$m(\Omega) = e^{-\eta|d(P_i,Q_i)-\beta|}, \tag{3.1c}$$

where P_i and Q_i are two matching points of the slices P and Q, $d(P_i, Q_i)$ is the corresponding distance; $\Omega = \{S_{PQ}, \overline{S_{PQ}}\}$, S_{PQ} means that points P_i and Q_i both belong to cortical osseous. Parameter β represents the tolerance that the expert associates with the value $d(P_i, Q_i)$, ρ is the inter-slice distance and η makes it possible to tolerate a greater inaccuracy in the geometrical resemblance of two consecutive contours. Dempster's rule was then used to combine the three mass functions for final segmentation.

In [Barhoumi et al., 2007], Barhoumi et al. introduced a new collaborative computer-aided diagnosis system for skin malignancy tracking. First, two diagnosis opinions were produced by perceptron neural network classification and content-based image retrieval (CBIR) schemes. Then Dempster's rule was used to fuse the two opinions to achieve a final malignancy segmentation. Simulation results showed that this BFT-based combination could generate accurate diagnosis results. In this work, the frame of discernment is composed of two elements, and the mass functions are Bayesian.

In [Guan et al., 2011], Guan et al. proposed a human brain tissue segmentation method with BFT. The authors first used Markov random forest (MRF) [Li, 2009] for spatial information segmentation and then a two-dimensional histogram (TDH) method of fuzzy clustering [Duan et al., 2008] to get a vague segmentation. Then a redundancy image was generated, representing the difference between the MRF and TDH methods, and Zhu's model (see Section 2.5.2.1) was used to calculate mass functions. Finally, Dempster's rule fused the two segmentation results and the generated redundancy image to handle controversial pixels. The visual segmentation results showed that this method has higher segmentation accuracy compared with the state-of-the-art.

As discussed in Section 2.2, Dempster's rule becomes numerically unstable when combining highly conflicting mass functions. In this case, the fused results can be unreliable, as a small change in mass functions can result in sharp changes in the fusion results. Researchers in the medical domain have also recognized this limitation. In [Ketout et al., 2012], Ketout et al. proposed a modified mass computation method to address this limitation and applied their proposal to endocardial contour detection. First, the outputs of each active contour model (ACM) [Kass et al., 1988] were represented as mass functions. Second, a threshold was proposed to check if the evidence m conflicts with others. If there was conflict, a modifying operation was used on the conflicting evidence. Finally, the results of edge set-based segmentation [Li et al., 2005] and region set-based segmentation [Chan et al., 2001] were fused by using the "improved BFT" [Xin et al., 2005] to get a more accurate contour of the left ventricle. Experimental results showed that the fused contour is closer to the ground truth than the contour from the edge or region. False detection of the two contours was suppressed in the resulting image by rejecting the conflicting events by the fusion algorithm. Meanwhile, the proposed method could detect the edges of the endocardial borders even in low-contrast regions.

In [Wen et al., 2018], Wen et al. proposed an improved MR image segmentation method based on FCM and BFT. First, the authors fused two modality images A and B with function $F(x,y) = w_1 g_A(x,y) + w_2 g_B(x,y)$, where x and y are image pixels and w_1 and w_2 are weighs used to adjust the influence of different images on the final fusion result and verifying $w_1 + w_2 = 1$. Second, the authors calculated the MV by FCM and calculated the mass functions of simple and double hypotheses by Zhu's model (see Section 2.5.2.1) without the neighboring pixel information. Third, the authors generated another kind of mass functions by weighting those of its neighboring pixels with the GD-based model (see Section 2.5.2.3) and used Zhu's model again to construct simple and double hypotheses mass functions. Finally, the authors used Dempster's rule to complete the fusion of the two mass functions to achieve the final segmentation. Compared with the FCM-based method, the proposed method can better decrease the conflict in multiple sources to achieve easy convergence and significant improvement by using Dempster's rule for classifier-level evidence fusion.

Besides, with the development of CNNs, the research community used Dempster's rule for the fusion of multiple CNN classifiers. In [George et al., 2020], George et al. proposed a breast cancer detection system using transfer learning and BFT. This first work first applied BFT in multiple evidence fusion with deep learning. High-level features were extracted using a convolutional neural network such as ResNet-18, ResNet-50, ResNet-101, GoogleNet, and AlexNet. A patch-based feature extraction strategy was used to avoid wrong segmentation of the boundaries and provide features with good discriminative power for classification. The extracted features were classified into benign and malignant classes using a support vector machine (SVM). A discounting operation was applied to transfer probability-based segmentation maps into mass functions. The discounted outputs from the different CNN-SVM frameworks were then combined using Dempster's rule. This work takes advantage of deep learning and BFT and has achieved good performance. Compared with majority voting-based fusion methods, BFT-based fusion showed superior segmentation accuracy. Compared with a single classifier, such as ResNet-101, the fused framework achieved an increase of 1%, 0.5%, 3%, and 2% for, respectively, sensitivity, specificity, and AUC. Also, the authors compared their results with the state-of-the-art method and achieved comparable segmentation accuracy.

Apart from using BFT to combine the discounted probabilities from the CNN classifiers [George et al., 2020], another solution is to construct a deep evidential segmentation framework directly. In [Huang et al., 2021a], We proposed a BFT-based semi-supervised learning framework (EUNet) for brain tumor segmentation. This work applied BFT in a deep neural network to quantify segmentation uncertainty directly. During training, two

Table 3.4: Summary of BFT-based medical image segmentation methods with multimodal inputs and several classifiers/clusterers.

Publications	Input image type	Application	BBA method
[Gautier et al., 2000]	multimodal MR images	lumbar sprain segmentation	Prior knowledge
[Lajili et al., 2018]	CT	breast segmentation	threshold
[Huang et al., 2021b]	PET/CT	lymphoma segmentation	None
[Huang et al., 2022a]	multimodal MR images	brain tumor segmentation	ENN

kinds of evidence were obtained: the segmentation probability functions and mass functions generated by UNet and EUNet, respectively. Dempster's rule was then used to fuse the two pieces of evidence. This is the first work that embeds BFT into CNN and achieves an end-to-end deep evidential segmentation model. We will introduce more detail about EUNet in Chapter 4.

The approaches introduced in this section use several classifiers or clusterers to generate different mass functions and fuse them by Dempster's rule. Among those approaches, George et al. [George et al., 2020] first applied Dempster's rule to combine the discounted probabilities from different deep segmentation models. In [Huang et al., 2021a], we merged ENN with UNet to construct an end-to-end segmentation model and fuse two kinds of evidence by Dempster's rule. Compared to George et al.'s approach [George et al., 2020], our approach can generate mass functions directly from a deep segmentation model, which is more promising.

3.2.2 *Multimodal evidence fusion*

Figure 3.7 shows an example of a medical image segmentation framework with several classifiers and multimodal inputs, which is more complex than the frameworks introduced in Sections 3.1.1, 3.1.2 and 3.2.1. The segmentation process comprises five steps: mass calculation, feature-level evidence fusion (optional), classifier-level evidence fusion, modality-level evidence fusion, and decision-making. Pixels at the same position from different modalities are fed into different classifiers and different BBA methods to get pixel-level mass functions. Dempster's rule is used first for feature-level evidence fusion, then to fuse classifier-level evidence, and last to fuse modality-level evidence. Here we show the segmentation example of the same located pixels z_i^1 and z_i^M that are obtained from modal 1 and M. The same pixel from different modalities is transferred separately into H classifiers, and some BBAs are used to get pixel-level mass functions. Similar to Figure 3.4, we assume we can obtain K mass functions with one classifier for each pixel. After the fusion of feature-level, classifier-level, and modality-level evidence fusion, a final mass function is assigned to the pixel z_i to represent the degree of belief this pixel belongs to class a, b and ignorance. Table 3.4 summarizes the segmentation methods focusing on classifier/clusterer fusion and modality-level evidence fusion.

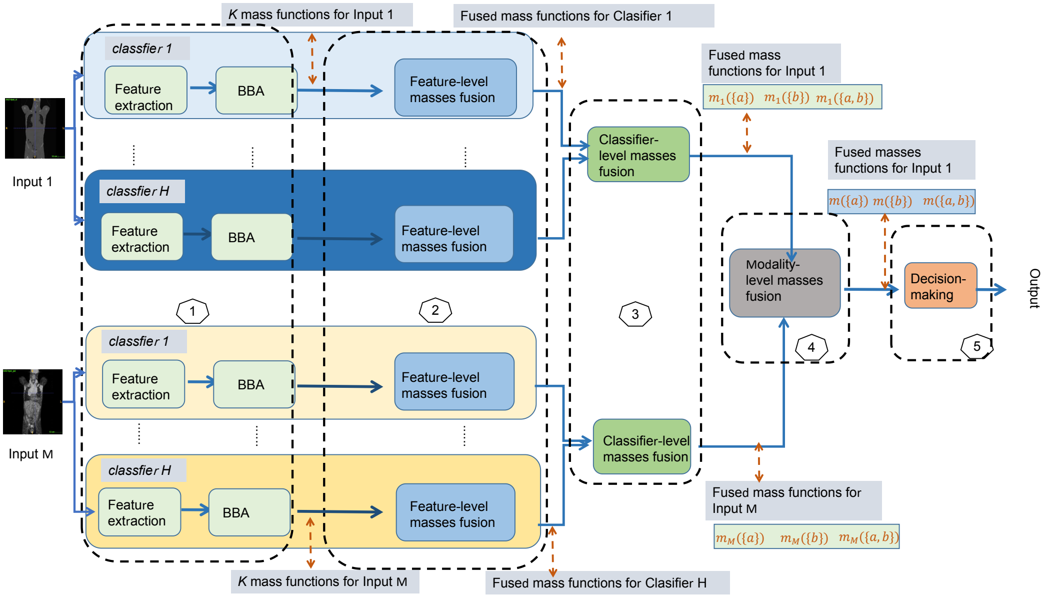

K mass functions for Input 1
Fused mass functions for Clasifier 1
classfier 1
Feature extraction
BBA
Feature-level masses fusion
Input 1
classfier H
Feature extraction
BBA
Feature-level masses fusion
Classifier-level masses fusion
Fused mass functions for Input 1
$m_1(\{a\})$ $m_1(\{b\})$ $m_1(\{a,b\})$
Fused masses functions for Input 1
$m(\{a\})$ $m(\{b\})$ $m(\{a,b\})$
Modality-level masses fusion
Decision-making
Output
1
2
3
4
5
classfier 1
Feature extraction
BBA
Feature-level masses fusion
Input M
classfier H
Feature extraction
BBA
Feature-level masses fusion
Classifier-level masses fusion
Fused mass functions for Input M
$m_M(\{a\})$ $m_M(\{b\})$ $m_M(\{a,b\})$
K mass functions for Input M
Fused mass functions for Clasifier H

Figure 3.7: Example framework of multimodal evidence fusion with several classifiers. The segmentation process is composed of five steps: (1) for each modality input, image features are fed into different classifiers, and some BBA methods are used to get feature-level mass functions; (2) inside each classifier, the feature-level mass functions are fused by Dempster's rule to get classifier-level mass functions; (3) inside each classifier, the calculated classifier-level mass functions are fused by Dempster's rule to get modality-level mass function; (4) inside each modal, the calculated modality-level mass functions are fused by Dempster's rule; (5) decision-making is made based on the final fused mass functions and outputs a segmented mask.

In [Gautier et al., 2000], Gautier et al. proposed a method for helping physicians monitor spinal column diseases with multimodal MR images. At first, an initial segmentation was applied with active contour [Lai, 1994]. Then several mass functions were obtained from expert opinions on constructing the frame of discernment. Thus, the mass functions were human-defined. Finally, Dempster's rule (see Section 2.2) was then used to fuse the mass functions from different experts. The method yielded the most reliable segmentation maps when the paper was published.

Based on their previous work on multimodal medical image fusion with a single cluster [Chaabane et al., 2009], Chaabane et al. presented another BFT-based segmentation method with several clusterers [Chaabane et al., 2011]. First, possibilistic C-means clustering [Bezdek, 2013] was used on R, G, and B channels to get three MVs. Then the MVs were mapped into mass functions with focal sets of cardinality 1 and 2 using Zhu's model (see Section 2.5.2.1). Dempster's rule was used first to fuse three mass functions from three corresponding color spaces. Based on the initial segmentation results, another mass function was calculated for each pixel and its neighboring pixels for each color space. Finally, the authors used Dempster's rule again to fuse the two mass functions from two corresponding clusterers. Experimental segmentation performance with cell images showed the effectiveness of the proposed method. Compared with the FCM-based segmentation method, the proposal increased by 15% the segmentation sensitivity.

In [Lajili et al., 2018], Lajili et al. proposed a two-step evidential fusion approach for breast region segmentation. The first evidential segmentation results were obtained by a gray-scale-based K-means clustering method, resulting in k classes. A sigmoid function was then used to define a mass function on the frame $\Omega = \{\text{Breast}, \text{Background}\}$ at each pixel z depending on its class. For local-homogeneity-based segmentation, the authors modeled the uncertainty with a threshold value α, by defining $m(\{\text{Breast}\}) = 1 - \alpha$, $m(\{\text{Background}\}) = 0$, $m(\Omega) = \alpha$, where α represents the belief mass corresponding to the uncertainty on the membership of the pixel z. A final fusion strategy with Dempster's rule was applied to combine evidence from the two mass functions. Experiments were conducted on two breast datasets [Suckling, 1994; Bowyer et al., 1996]. The proposed segmentation approach yielded more accurate results than the best-performed method. It extracted the breast region with correctness equal to 0.97, which was 9% higher than the best-performing method.

In [Huang et al., 2021b], we proposed to consider PET and CT as two modalities and used to UNet model to segment lymphoma separately. Then the two segmentation masks were fused by Dempster's rule. Although this is the first work that applied BFT in multimodal evidence fusion with several deep segmentation models, a limitation of this work is that only Bayesian mass functions are used for evidence fusion. To improve this first work, we proposed a multimodal evidence fusion model with contextual discounting for brain tumor segmentation [Huang et al., 2022a]. In this work, using four modules for feature extraction and evidential segmentation, the framework assigns each voxel a mass function. In addition, a contextual discounting layer is designed to take into account the reliability of each source when classifying different classes. Finally, Dempster's rule is used to combine the discounted evidence to obtain a final segmentation. This method can be used together with any state-of-the-art segmentation module to improve the final performance. We will introduce more detail about the multimodal evidence fusion model in Chapter 6.

Few studies have considered multimodal medical images as independent inputs and used independent classifiers to generate mass functions. The performance of this kind of approach is limited by the representation of image features and the ability to quantify the model uncertainty. We first merged ENN with UNet for the fusion of multimodal MR images with contextual discounting [Huang et al., 2022a]. This approach enables the model to generate a learned reliability metric from input modalities during different segmentation tasks, which can potentially make the results more explainable.

3.3 CONCLUSION

The choice of using single-modality or multimodal depends on the dataset. Generally, the more source data we have, the more reliable segmentation results we will get. The choice of a single or several classifiers/clusterers depends on the limitation of the computation source and the requirement of computation efficiency. Prior to 2020, BFT-based medical image segmentation methods had limited segmentation accuracy due to the use of low-level image features, such as grayscale and shape features, to generate mass functions. Moreover, none of them considered the segmentation reliability. Since the application of deep learning in medical image segmentation has been very successful, the use of BFT in deep neural networks is a promising research direction, in particular, to quantify the uncertainty and reliability of the segmentation results, as well as fuse evidence from multimodal medical images. In the next part, we will introduce our contributions related to the construction of the deep segmentation model with BFT and deep learning algorithms (Chapter 4), the uncertainty quantification with deep segmentation models (Chapter 5), and multimodal evidence fusion (Chapter 6).

Part II

DEEP EVIDENTIAL SEGMENTATION

Chapter 4

Semi-supervised medical image segmentation

Contents

4.1 Introduction

The performance of deep learning-based methods depends mainly on large-scale learning examples and their annotations. However, obtaining precisely annotated data is very challenging in the medical domain. Recently, some semi-supervised deep learning methods have been proposed to train the models with fewer labels. Techniques for semi-supervised medical image segmentation can be divided into three classes: graph-constrained methods [Xu et al., 2017], self-learning methods [Li et al., 2019a; Min et al., 2019], and generative adversarial learning methods [Mondal et al., 2018; Sun et al., 2019]. Though experimental results are promising, only a few authors focus on studying the uncertainty caused by the low quality of the images and the lack of annotations. In this work, we follow the main idea of self-training and use the image information to construct a semi-supervised brain tumor segmentation framework. We first propose to use two parallel segmentation modules: a probabilistic segmentation module and an evidential segmentation module, to obtain two segmentation results. We then use an evidential fusion module to combine evidence in order to decrease uncertainty. Results from a series of experiments on the BraTS2019 brain tumor dataset showed that our framework achieves encouraging results when only some training data are labeled.

We organize this chapter as follows: Section 4.2 introduces the semi-supervised medical image segmentation framework. Section 4.3 reports numerical experiments by comparing the segmentation performance with the state-of-the-art methods under supervised and semi-supervised learning. Finally, we conclude this work in Section 4.4.

4.2 PROPOSED FRAMEWORK

In this section, we describe the proposed framework, named SEFNet, which is composed of a feature extraction module similar to UNet, a probability assignment module to obtain a probability segmentation, a basic belief assignment module to obtain an evidential segmentation, and an evidence fusion module to combine both probability and evidential segmentations. Figure 4.1 shows the overall flowchart of our proposal. The first step is to extract features from input images. Second, two modules are used in parallel to map features into probabilities or mass functions. The probability assignment module uses the softmax transformation to map the extracted features into probabilities. The mass function assignment module uses ENN (see Section 2.5.1.2) to map the extracted features into mass functions. Third, the two sources of evidence, probabilities and mass functions, are combined by Dempster's rule in the evidence fusion module. The semi-supervised learning algorithm will be introduced in Section 4.2.2.

4.2.1 *Evidential segmentation with multiple evidence fusion*

Figure 4.2 shows the detailed evidential segmentation framework. The input is composed of four MRI modalities. (Here, we show an example of input data of size $4 \times 128 \times 128 \times 128$.)

FEATURE EXTRACTION As shown in Figure 4.2, we apply multi-fiber units (see Figure 1.9 from Section 1.3.2) in the encoding stage to achieve multi-scale representation. In the decoding stage, the high-resolution features from the encoding stage are concatenated with the upsampled features, which makes the whole feature extraction module similar to UNet [Ronneberger et al., 2015].

PROBABILITY AND BELIEF ASSIGNMENT MODULES The *probability assignment* module comprises a $1 \times 1 \times 1$ projection layer followed by a softmax layer, which maps the feature vectors into probabilities directly. The

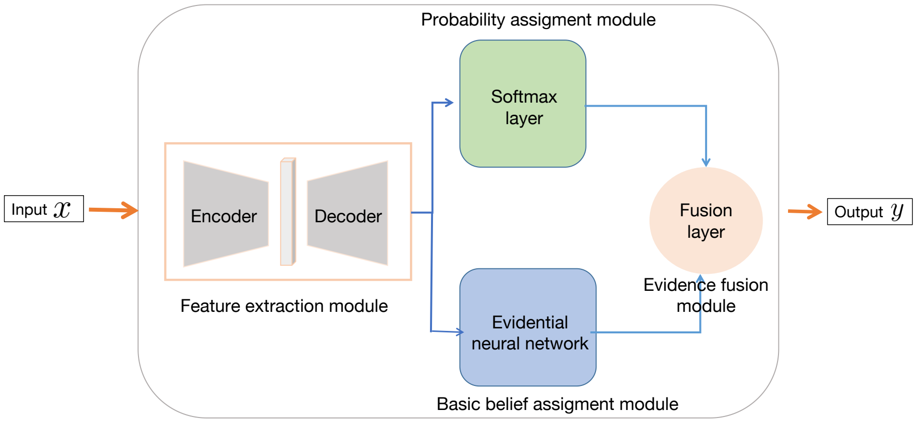

Probability assigment module
Softmax layer
Input x
Encoder
Decoder
Feature extraction module
Fusion layer
Output y
Evidence fusion module
Evidential neural network
Basic belief assigment module

Figure 4.1: Overall flowchart of our proposal. It is composed of four modules for feature extraction, probability assignment, basic belief assignment, and evidence fusion.

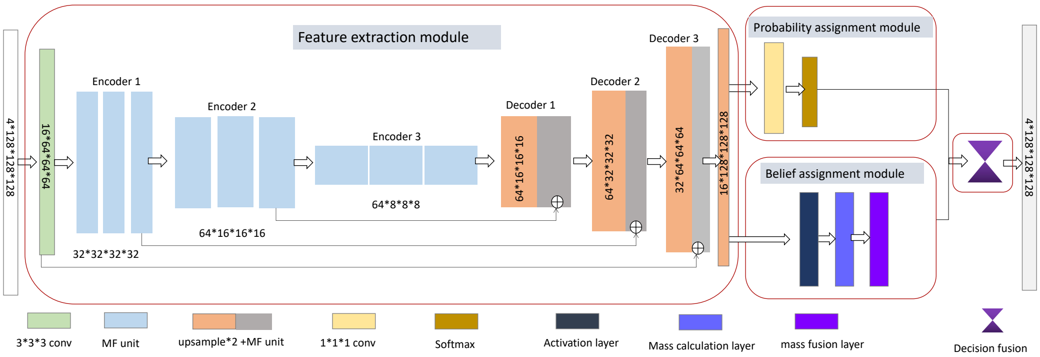


Figure 4.2: Proposed evidential medical image segmentation framework.

output of the probability assignment module is denoted as p_{CNN}, and we assume that each voxel belongs to one of four classes denoted as $\{0, 1, 2, 4\}$. The *belief assignment* module is based on the ENN model recalled in Section 2.5.1.2; it comprises three layers: an prototype activation layer that computes distance-based activations using (2.19), a mass calculation layer that computes mass functions using (2.20), and a combination layer that combined mass functions derived from prototypes using Dempster's rule.

EVIDENCE FUSION The objective of this module is to make a final segmentation decision. The decision based on several information sources can be expected to be more accurate and reliable than using a single source of information. In our case, if only part of the training data is labeled, the uncertainty is higher than it is in the fully supervised case. To increase the segmentation performance, we propose an additional *evidence fusion* layer to combine evidence from the probability and belief assignment modules. Here, the voxel-wise output probability distributions p_{CNN} from the probability assignment module can be seen as a Bayesian mass functions, which can be combined with the voxel-wise output mass functions m_{ENN} from the belief assignment module using Dempster's rule. The combined mass functions are Bayesian, and are given by

$$(p_{\mathrm{CNN}} \oplus m_{\mathrm{ENN}})(\{\omega_c\}) = \frac{p_{\mathrm{CNN}}(\omega_c) pl_{\mathrm{ENN}}(\omega_c)}{\sum_{l=1}^{C} p_{\mathrm{CNN}}(\omega_l) pl_{\mathrm{ENN}}(\omega_l)}, \quad c = 1, \ldots, C, \tag{4.1}$$

where $pl_{\mathrm{ENN}}(\omega_c) = m_{\mathrm{ENN}}(\{\omega_c\}) + m_{\mathrm{ENN}}(\Omega)$ is the plausibility of class ω_c derived from mass function m_{ENN}.

4.2.2 *Semi-supervised learning*

We propose a semi-supervised learning algorithm to optimize the framework when only part of the training data is labeled, with the aim of obtaining an accuracy as close as possible to that of a fully supervised learning method. The general idea is that similar images are expected to produce similar classification or segmentation results even if some transformations have been performed because the relevant characteristics are preserved despite the transformation. During each learning epoch, a transformed copy x_t of each input image x is computed using one of several transformations, namely, random intensity change, Gaussian blur and exponential noise. This transformation operation will be described in Section 4.3.1.2. Two loss functions are proposed for training data with and without labels.

TRAINING WITH LABELS We train the network with the labeled data using the following $\mathsf{loss1}$ function, class-independent Dice loss, which measures the overlap region between the output S and the ground truth G:

$$\mathsf{loss1} = \sum_{c=1}^{C}\left(1 - 2\frac{\sum_{n=1}^{N_1} S_{cn} G_{cn}}{\sum_{n=1}^{N_1} S_{cn} + G_{cn}}\right), \tag{4.2}$$

where $G_{cn} = 1$ if voxel n belongs to class c, and $G_{cn} = 0$ otherwise, and S_{cn} represents the corresponding predicted mask; C is the number of classes and N_1 is the number of voxels with labels.

TRAINING WITHOUT LABELS For the data without labels, we use the mean squared error loss, denoted here as $\mathsf{loss2}$, to optimize the feature representation by minimizing the difference between the original output S and the transformed output S_t:

$$\mathsf{loss2} = \frac{1}{2N_2C}\sum_{c=1}^{C}\sum_{n=1}^{N} \left\|S_{cn} - S_{cn}^{t}\right\|^2, \tag{4.3}$$

where S_{cn}^{t} is the segmented label at point n corresponding to the generated input x_t, N_2 is the number of voxels without labels.

4.3 EXPERIMENTS AND RESULTS

In this section, we present numerical experiments to verify the effectiveness of the proposed model. In Section 4.3.1, we introduce the dataset, the preprocessing of the data, the parameter setting, and the evaluation protocols. The sensitivity analysis of hyper-parameters and the comparative analysis of segmentation performance are then introduced in Sections 4.3.2 and 4.3.3, respectively.

4.3.1 *Experiment settings*

4.3.1.1 *Dataset*

The experiment data is provided by the Brain Tumor Segmentation (BraTS) 2019 challenge [Menze et al., 2014; Bakas et al., 2017; Bakas et al., 2018]. The dataset consists of 335 cases of patients for training, 125 cases for validation and 166 cases for test. Since the test set is not available now, in this chapter, we use the official validation set to test our model, i.e., take it as a test set. For each patient, we have four kinds of MR sequences: T1, T1-weighted (T1Gd), T2, and FLAIR. Each of them has a volume of $155 \times 240 \times 240$. For data from the training set, each case was annotated into three heterogeneous histological sub-regions: peritumoral edema (ED, label 2), necrotic core and non-enhancing tumor (NRC/NET, label 1), and enhancing tumor(ET, label 4). The background is marked as label 0. Figure 4.3 shows an example of the four modalities and the corresponding tumor region. The evaluation was based on the segmentation accuracy of

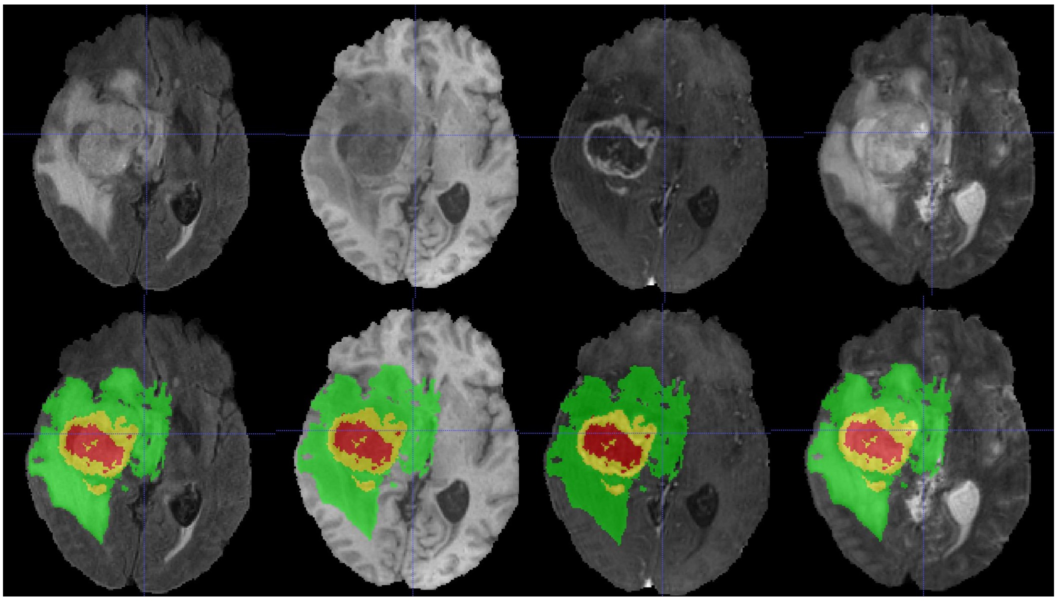


Figure 4.3: Example of one patient from BraTs2019 dataset. The first and second rows show MR images and MR images with labeled tumor masks, respectively. From left to right: FLAIR, T1, T1Gd, T2. Labels 1, 2, and 4 are marked in red, green, and yellow, respectively.

three overlap regions: enhancing tumor (ET, label 4), tumor core (TC, the composition of label 1 and 4), and whole tumor (WT, the composition of label 1, 2, and 4). For data from the validation set, only four modalities of MR sequence information are available.

We used five-fold cross-validation to train our SEFNet. During training, we randomly divided the BraTS 2019 training set into 5 equal-sized datasets. The training process was then repeated five times, with each of the 5 datasets used exactly once as the validation data. The segmentation performance with cross-validation is reported by the average of five models. For a fair comparison with the state-of-the-art, we fine-tuned the best-performing model with the full training set and tested the performance on the validation set. The segmentation performances were assessed by the online evaluation server CBICA's Image Processing Portal [1].

4.3.1.2 *Pre-processing*

Before feeding the data into the framework, several pre-processing methods were used to process the input data. We first applied intensity normalization to each MRI modality from each patient independently by subtracting the mean and dividing by the standard deviation of the brain region only. Moreover, to prevent overfitting, we used four types of data augmentation. First, we applied a random intensity shift between $[-0.1, 0.1]$ and random intensity scaling between $[0.9, 1.1]$ to MRI data. Second, we

[1] https://ipp.cbica.upenn.edu/

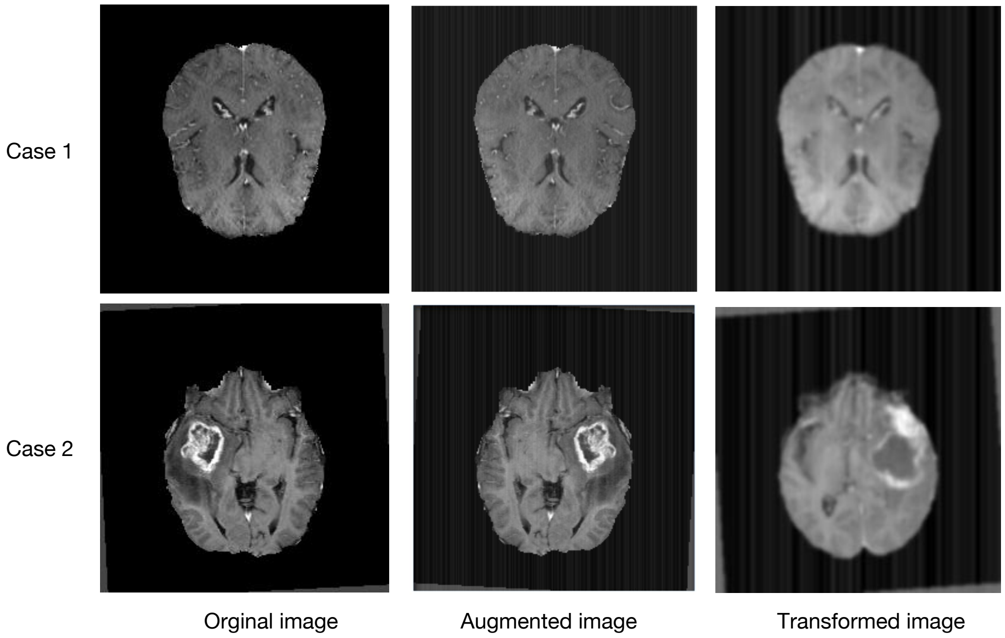


Figure 4.4: Example of pre-processed images. From left to right: the original image, the augmented image and the transformed image, respectively.

randomly cropped the MRI data from $155 \times 240 \times 240$ to $128 \times 128 \times 128$. Third, we used random rotation with a rotation angle of 10. Finally, we used random mirror flipping for MRI data along each 3D axis with a flip probability of 50%. The data augmentation operation is applied during each training epoch. Since the data augmentation operation is randomly chosen, the input x in each training epoch varies.

For semi-supervised training, we used transformations for each preprocessed input x. We first applied random intensity change on the input with the shift between $[-0.2, 0.2]$ and scaling between $[0.9, 1.1]$. Then we added Gaussian Blur with a standard deviation of 3 to the image. Finally, we added exponential noise with an exponent of 3 to the input. After transformation, for each input x, we will generate transformed data x_t. Figure 4.4 shows two examples of input images before and after prepossessing. (To better show the difference between images, here we only show the aligned images without random cropping). Compared with the original image, the augmented image is randomly flipped with a small intensity change and image rotation. Compared with augmented input x, the transformed input x_t is more blurred and noisy.

4.3.1.3 *Parameter settings*

For the feature extraction module, the spatial dimension and input channel were set, respectively, as 3 and 4. The channels (number of filters) of the input layer were set as 16. The channels (number of filters) of MF units were set as $32, 64, 64$ for three corresponding encoders and each MF unit has 16 parallel fibers (see Section 1.3.2).

For the mass assignment module, the prototypes p (2.19) were initialized randomly with Xavier uniform [Hanin et al., 2018] distributions to simplify the model, as well as the parameter of the membership degrees u (2.20). The impact of prototype numbers will be discussed in Section 4.3.2.1. Then we initialized parameters α (2.19) and γ (2.19) with constants 0.5 and 0.01.

The maximum training epoch was set to 300. The training process was stopped when the performance did not increase in 20 epochs. The Adam optimization algorithm with batch size 8 was used to train the model. The initial learning rate was set to 0.001 at the beginning and decayed with an adjusted learning rate

$$lr = lr_0 \times \left(1 - \frac{e}{N_e}\right)^{0.9}, \tag{4.4}$$

where e is an epoch counter, and N_e is a total number of epochs. The model with the best performance with cross-validation was saved as the final model for testing. All experiments were implemented in Python with the PyTorch framework and were trained and tested on a desktop with a 2.20GHz Intel(R) Xeon(R) CPU E5-2698 v4 and a Tesla V100-SXM2 graphics card with 32 GB GPU memory.

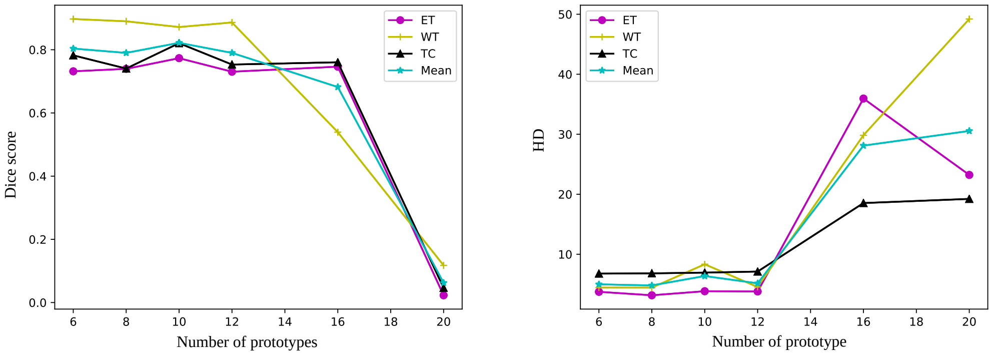
4.3.1.4 *Evaluation criteria*

We used two evaluation criteria: the Dice score (see (1.1)) and the Hausdorff Distance (HD) (see (1.4)), to measure the segmentation performance. For each patient, we separately computed these two indices for the three classes and then averaged indices over the patients. Results are reported in Sections 4.3.2 and 4.3.3.

4.3.2 *Sensitivity analysis*

4.3.2.1 *Evaluating the impact of number of prototypes*

The number of prototypes is an important hyper-parameter that may impact segmentation performance. We trained the model with 6, 8, 10, 12, 16, and 20 prototypes. Figure 4.5 shows the corresponding results. We can see from Figure 4.5 when the number of prototypes is between 6 and 12, the performance is stable. SEFNet achieves the best performance in terms of Dice score and Hausdorff Distance with 10 prototypes. With more than 12 prototypes, the performance decreases. Thus, in the following experiments, the number of prototypes in SEFNet is fixed as 10.



(a)

(b)

Figure 4.5: Values of the (a) Dice score (the higher the better), (b) Hausdorff distance (HD) (the lower the better) of ET, TC, WT with the different number of prototypes.

4.3.2.2 *Evaluating the impact of semi-supervised learning*

To evaluate the impact of the proportions of the labeled training data, we trained our model with 100%, 70%, 50%, and 30% labeled data and reported the cross-validation performance in Table 4.1. It shows the corresponding segmentation performance of the Dice score and the Hausdorff distance (HD), as well as the corresponding standard deviations (SD) among the evaluated cases. Compared with 100% training data labeled, the Dice score decreased by 2.1%, 5.6%, and 7%, respectively, when only 70%, 50%, and 30% training data are labeled. With different proportions of the labeled training data, the model shows comparable performance on Hausdorff Distance. Compared with WT (ET+NRC/NET+ED) segmentation, the standard deviations (SD) are quite large for ET and TC (ET+NRC/NET) segmentation. Two main reasons can explain the large SD: the segmentation between ET and NRC/NET is challenging; some cases do not contain ET or NRC/NET tumors, while the model segment DE tumors into ET or NRC/NET class. Figure 4.6 shows plots of the Dice loss and Dice score under different proportions of training labels during training. With the decrease in training labels, the model only slightly decreases segmentation performance. Table 4.2 shows the segmentation performance on the online validation dataset with the five corresponding models obtained by cross-validation. When only 70%, 50%, and 30% training data are labeled, the Dice score decreased by 2.9%, 4.2%, and 7.5%, respectively, compared with 100% training data labeled. The above results demonstrate the effectiveness of semi-supervised learning.

Furthermore, we compared the distribution of the Dice score among 125 validation cases in our model SEFNet and the baseline model MFNet, under the different proportions of labeled training data in Figure 4.7. To

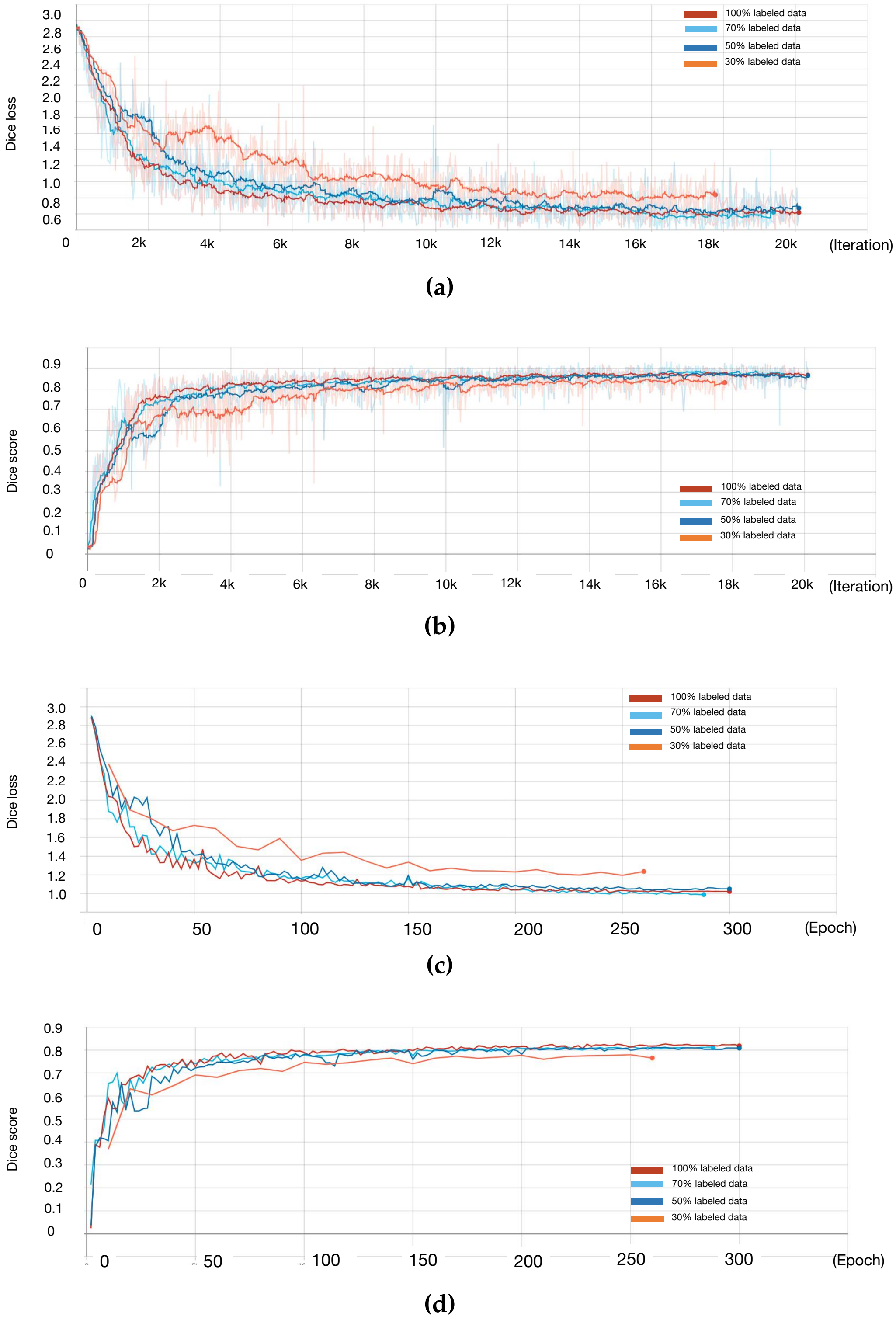


Figure 4.6: Plots of Dice loss and Dice score during training. (a) Mean Dice loss during training, (b) Mean Dice score during training, (c) Mean Dice loss with cross-validation, (d) Mean Dice score with cross-validation.

Table 4.1: Performance comparison on the BraTS2019 training set with cross-validation.

Label	Dice Score(±SD)			HD(±SD)		
proportion	ET	WT	TC	ET	WT	TC
100 %	0.783±0.227	0.906±0.056	0.805±0.220	3.481±4.516	4.618±3.829	6.978±7.595
70 %	0.762±0.215	0.905±0.056	0.814±0.201	5.151±7.254	6.219±11.037	6.507±7.732
50 %	0.727±0.261	0.904±0.066	0.806±0.215	5.115±6.149	4.858±5.399	6.672±7.529
30 %	0.713±0.247	0.897±0.070	0.799±0.215	5.904±7.214	5.554±6.867	6.459±6.489

Table 4.2: Performance comparison on the BraTS2019 online validation set.

Label	Dice Score (±SD)			HD(±SD)		
proportion	ET	WT	TC	ET	WT	TC
100%	0.763±0.254	0.883±0.095	0.808±0.201	4.592±8.401	5.983±6.202	7.671±10.608
70%	0.734±0.286	0.884±0.110	0.767±0.242	5.021±8.929	5.718±7.929	8.427+11.799
50%	0.721±0.294	0.877±0.129	0.777±0.227	7.204±17.002	7.942±15.235	11.830±19.809
30%	0.688±0.289	0.887±0.113	0.744±0.267	8.609±17.533	7.448±12.585	12.236±19.690

simplify the comparison, we only showed the mean value of the Dice score here. The boxplot displays the data based on a five-number summary: the minimum (the lowest data point excluding any outliers), the maximum (the largest data point excluding any outliers), the median, and the 25% and 75% percentile. As shown in Figure 4.7, SEFNet yields better performance than MFNet with different proportions of labeled training data. When 100% training data are labeled, SEFNet and MFNet can achieve high segmentation accuracy. With the decreasing proportion of labeled training data, the SEFNet shows increasing advantages. Compared with MFNet, when only 30% training data are labeled, SEFNet yields around 4%, 5%, and 8% increase, respectively, in ET, WT, and TC.

4.3.3 *Comparative analysis*

4.3.3.1 *Comparison with full-supervised methods*

We first compare our results with the state-of-the-art under full-supervised learning on the BraTs2019 validation set. The comparison is presented in Table 4.3. We highlight the best-performed results in bold characters and underline the second-best results. SEFNet achieves a Dice score of 0.793, 0.868, 0.861, and 0.841, respectively, for ET, WT, TC, and the mean over the three regions. Compared with MFNet, it has an increase of 4%, 6.2%, and 4.2% of Dice score in ET, TC and the mean, respectively. Also, SEFNet surpasses most of the reported methods, i.e., UNet, attention UNet, and MCNet. The performance of SEFNet is not as good as the top one solution of the BraTs2019 challenge segmentation task, which uses a Two-stage cascaded U-Net [Jiang et al., 2019] (double model), with marginal performance gaps of 0.9%, 4.1%, and 0.3% for ET, WT, and TC in Dice score, respectively. The two-stage cascaded U-Net framework uses one UNet

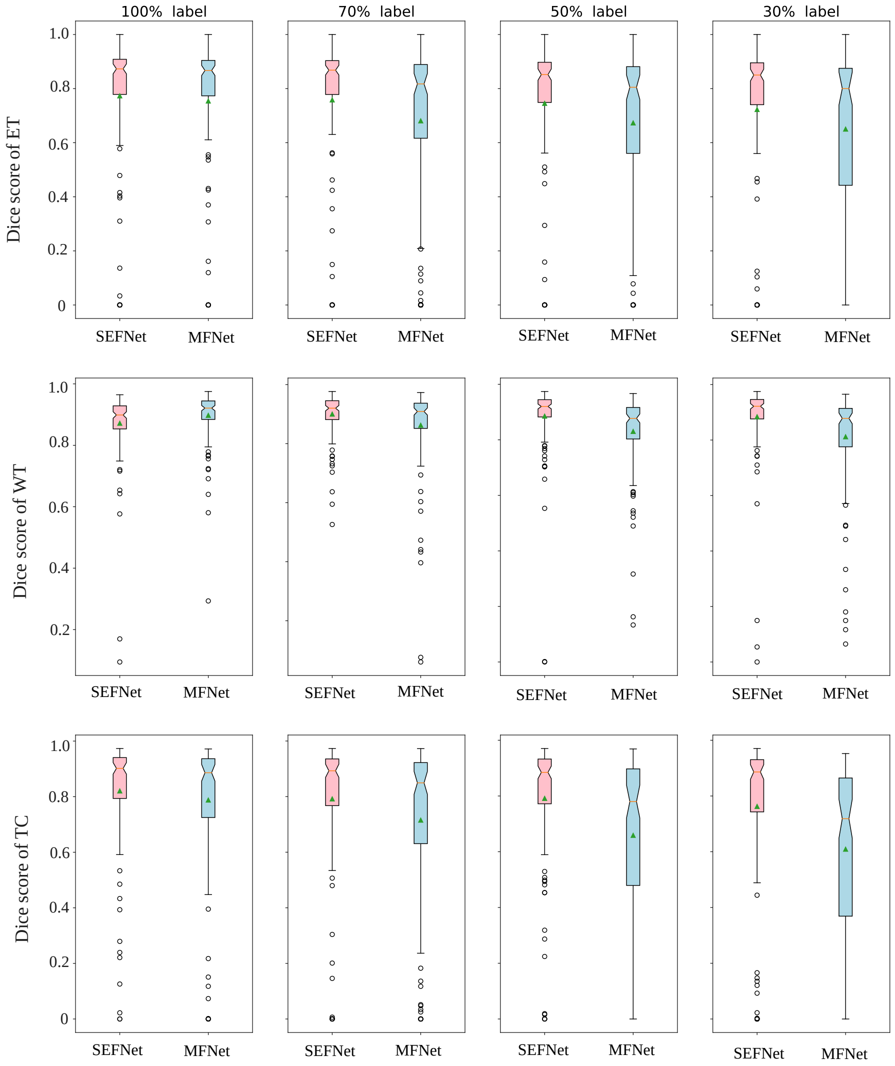


Figure 4.7: Dice score of ET, TC, and WT with different percentages of labeled training data. From left to right: the results of training with 100%, 70%, 50%, and 30% labels, respectively. The first, second, and third rows show the Dice score of ET, WT, and TC, respectively. The pink and light-blue boxplot represents the results of our proposal (SEFNet) and the baseline model (MFNet), respectively. The orange line and green triangle represent the median and mean value of the Dice score, respectively.

Table 4.3: Performance comparison on the BraTs2019 validation set (in case of full-supervised learning). The best results are in bold, and the second best results are underlined.

Methods	Dice score				HD			
	ET	WT	TC	Mean	ET	WT	TC	Mean
SEFNet (ours)	0.793	0.868	0.861	0.841	5.616	8.329	6.618	6.854
MFNet [Chen et al., 2019a]	0.753	0.880	0.765	0.799	4.872	8.022	9.706	7.562
DMFNet [Chen et al., 2019a]	0.756	0.890	0.799	0.815	5.069	6.531	7.454	6.351
3D-UNet [Wang et al., 2019]	0.737	0.894	0.807	0.812	5.994	5.567	7.357	6.342
Dense-UNeT[Agravat et al., 2019]	0.600	0.700	0.630	0.643	11.690	14.330	17.100	14.373
AttentionUNet [Islam et al., 2019]	0.704	0.898	0.792	0.798	7.050	6.290	8.760	7.370
MCNet [Li et al., 2019b]	0.771	0.886	0.813	0.823	6.232	7.409	6.033	6.558
Two-stage cascaded U-Net [Jiang et al., 2019]	**0.802**	**0.909**	**0.864**	**0.858**	**3.145**	**4.263**	**5.439**	**4.282**

for coarse segmentation and then another UNet for accurate segmentation, which could increase the segmentation accuracy to a certain degree. However, the computation cost is high and the reported memory requirement is over 12 GB during the experiment with a batch size of 1.

Figure 4.8 presents a visual comparison of the brain tumor segmentation results obtained from different slices. From left to right, we can see the ground truth (GT), the segmentation result of the baseline method (MFNet) and our proposal (SEFNet), and the **difference map** between MFNet and GT on the one hand and between SEFNet and GT on the other hand. The white points in the **difference map** indicate the position the voxels are wrong-segmented. We highlight the regions with fewer misclassification voxels by red circles in Figure 4.8, where there are fewer white points. Compared with MFNet, SEFNet can generate more precise segmentation results.

4.3.3.2 *Comparison with semi-supervised methods*

We also compared the segmentation performance with the state-of-the-art under semi-supervised learning. Comparing our result with those of other semi-supervised methods is difficult because BraTs datasets from different years (2015, 2017, 2018, or 2019) were tested with those methods, leading to the various training and validation set components. Also, the different propositions of training labels make the comparison difficult. In this work, we simplify the comparison by only comparing the results when 50% of the training data are labeled. As we can see from Table 4.4, SEFNet achieves the best performance on validation data with the reported 0.792 mean Dice score on BraTS2019 dataset. Compared with the performance of MASSL and SAM-GAN on the BraTs2018 dataset, SEFNet yields an increase of 10% and 4.1% in the Dice score of WT and Mean, respectively. Compared with the best performance of PGAN on the BraTs2017 dataset, SEFNet has an increases of 16.6%, 12.8%, and 8.9% in the Dice score of WT, TC, and Mean, respectively. Compared with the performance of TSMAN on the BraTs2015

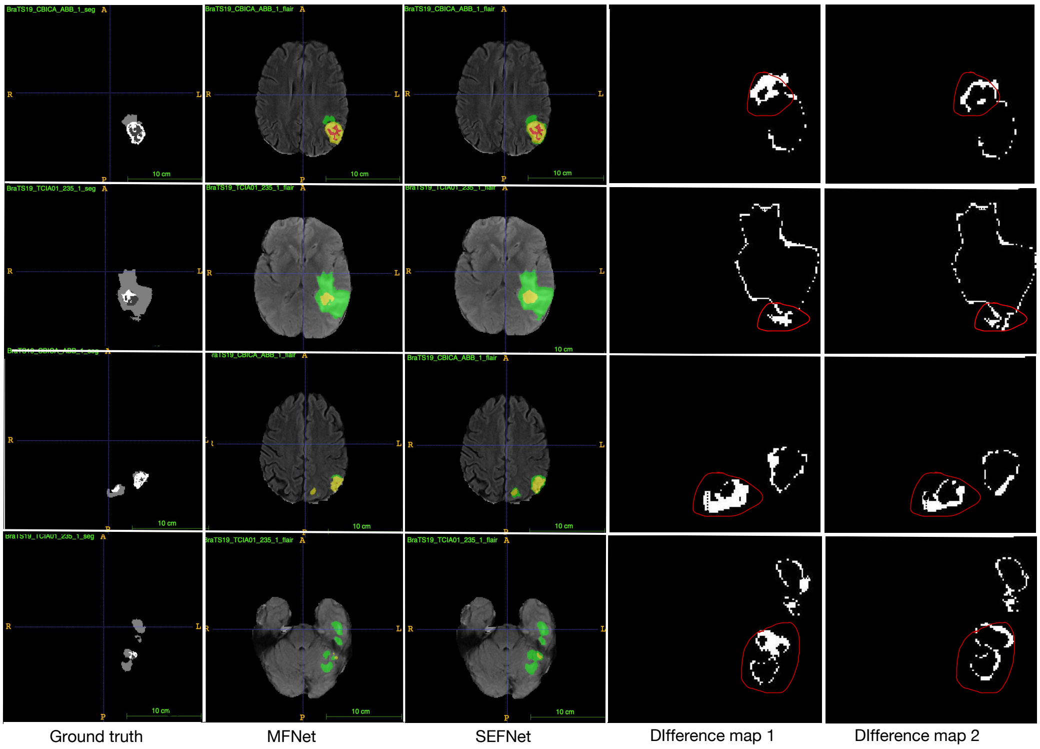


Figure 4.8: Example segmentation results on the BRATS2019 training dataset. The figures in rows show the results of different axial slices from left to right: the ground truth (GT), the segmentation results of MFNet and SEFNet, and the **difference map** between GT and the corresponding segmentation map. The points in white indicate where the segmented result is wrong. Labels 1, 2, and 4 are marked in red, green, and yellow, respectively. Moreover, we highlight the regions with fewer misclassification voxels by red circles in **difference map**.

Table 4.4: Performance comparison on the BraTs2019 validation set (in case of semi-supervised learning). Symbol $*$ indicates the results are not available from the published paper.

Dataset	Method	Training	Test	Proportion of	Dice score			
		Number	Number	labeled data	ET	WT	TC	Mean
BraTs2019	SEFNet (ours)	285	125	50%	0.721	**0.877**	**0.777**	**0.792**
BraTs2018	MASSL [Chen et al., 2019b]	200	50	50%	*	0.770	*	*
	SAM-GAN[Xi, 2019]	285	66		*	*	*	0.751
BraTs2017	Transfer-UNet[Zeng et al., 2017]	285	46	50%	0.734	0.690	0.631	0.685
	PGAN [Sun et al., 2019]	285	46		**0.751**	0.711	0.649	0.703
BraTs2015	Transfer-UNet [Zeng et al., 2017]	140	80	50%	0.633	0.616	0.642	0.630
	PGAN [Sun et al., 2019]	140	80		0.668	0.652	0.667	0.662
	TSMAN [Min et al., 2019]	244	30		*	*	*	0.707

dataset, SEFNet has an increase of 9% in the mean Dice score. The above comparison demonstrates the effectiveness of SEFNet.

4.4 CONCLUSION

In this chapter, we have presented a semi-supervised multiple evidence fusion framework (SEFNet) for medical image segmentation. With the SEFNet framework, we compute two pieces of segmentation evidence: probability functions generated by a softmax layer, and mass functions generated by an ENN module. Dempster's rule is then used to fuse the two pieces of evidence and to decrease segmentation uncertainty. For images with labels, we use the supervised class-independent Dice loss to guide the training process. For images without labels, we use information constraints through image transformation operations to provide training guidance. Quantitative and qualitative results on the BraTs2019 dataset show that using Dempster's rule with semi-supervised learning makes it possible to efficiently deal with segmentation uncertainty, resulting in comparable performance under semi-supervised conditions.

There are some limitations to this work. First, the evidence from the basic belief assignment module needs to be well studied; in particular, the study of masses in the frame of discernment could be interesting and meaningful. Second, one of the potential problems in this work may be the independence of sources of information. The two pieces of evidence obtained from the probability assignment module and the basic belief assignment module are not independent because they share the same features. In the next chapter, we will first study the segmentation uncertainty and then further evaluate the framework by applying it to other medical image segmentation problems. Lastly, we will investigate the reliability of segmentation results with expected calibration error (ECE).

Chapter 5

Uncertainty quantification in medical image segmentation

Contents

5.1 introduction

Despite the excellent performance of deep learning methods, the issue of quantifying prediction uncertainty remains [Hüllermeier et al., 2021]. This uncertainty can be classified into three types: distribution, model, and data uncertainty. Distribution uncertainty is caused by training-test distribution mismatch (dataset shift) [Quinonero-Candela et al., 2008]. Model uncertainty arises from limited training set size and model misspecification [Mehta et al., 2019; Maddox et al., 2019; Yu et al., 2019]. Finally, sources of data uncertainty include class overlap, label noise, and homo or hetero-scedastic noise. Because of the limitations of medical imaging and labeling technology, as well as the need to use a large nonlinear parametric segmentation model, medical image segmentation results are particularly tainted with uncertainty, which limits the segmentation reliability.

In clinical lymphoma diagnosis and radiotherapy planning, PET-CT scanning is an effective imaging tool to locate and segment lymphomas. The standardized uptake value (SUV), defined as the measured activity

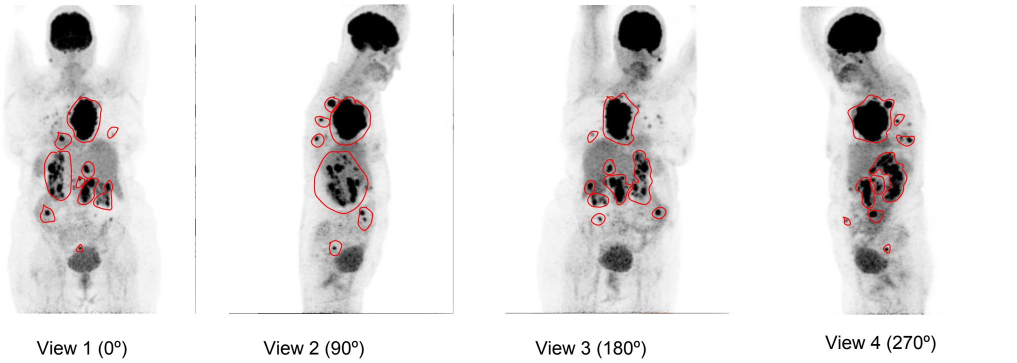


Figure 5.1: Example of a patient with lymphomas in 3D PET-CT views. The lymphomas are the dark areas circled in red.

normalized for body weight and injected dose to remove variability in image intensity between patients, is widely used to locate and segment lymphomas thanks to its high sensitivity and specificity to the metabolic activity of tumors [Jhanwar et al., 2006]. However, PET images have a low resolution and suffer from the partial volume effect blurring the contours of objects [Zaidi et al., 2010]. For that reason, CT images are usually used jointly with PET images because of their anatomical feature-representation capability and high resolution. Figure 5.1 shows 3D PET-CT views of a lymphoma patient in which PET and CT are merged for showing. The lymphomas are in black, as well as the brain and the bladder. As we can see from this figure, lymphomas vary in intensity distribution, shape, type, and number. The attributes of lymphomas make the study of automatic lymphoma segmentation challenging and unreliable.

In this chapter, we introduce a different uncertainty quantification approach in medical image segmentation. The main idea is to hybridize a deep medical image segmentation model with evidential classifiers (one of the classifiers, ENN, was introduced in Section 2.5.1.2, and the other is the Radial basis function (RBF) network-based evidential classifier). Two evidential layers are first compared based on different ways of using distances to prototypes for computing mass functions, to show the effectiveness of uncertainty quantification. Then the two evidential classifiers are plugged with deep segmentation models to construct deep evidential segmentation models to quantify segmentation uncertainty. The whole model is trained end-to-end by minimizing the Dice loss function. The proposed combination of deep feature extraction and evidential segmentation is shown to outperform the baseline model as well as three other state-of-the-art models on a dataset of 173 patients with lymphomas.

We organize this chapter as follows: Section 5.2 first compares the performance of uncertainty quantification of two evidential classifiers and then introduces two deep evidential segmentation models with the application on lymphoma segmentation. Section 5.3 reports the numerical experiments, which demonstrate the advantages of uncertainty quantification of the proposed segmentation models. Finally, we conclude this work in Section 5.4.

5.2 PROPOSED METHOD

5.2.1 *Uncertainty quantification with evidential classifiers*

In this section, we review two methods for designing classifiers that output mass functions to quantify uncertainty, referred to as *evidential classifiers*. The Evidential neural network (ENN) classifier introduced in [Denœux, 2000] is first recalled in Section 2.5.1.2. A new model based on the interpretation of a radial basis function (RBF) network as combining of simple mass functions by Dempster's rule, inspired by [Denœux, 2019b], is then described in Section 5.2.1.2. The two models are compared experimentally in Section 5.2.1.3.

5.2.1.1 *Evidential neural network (ENN)*

In [Denœux, 2000], Denœux proposed the ENN classifier, in which mass functions are computed based on distances to prototypes. The basic idea is to consider each prototype as a piece of evidence, which is discounted based on its distance to the input vector. The evidence from different prototypes is then pooled by Dempster's rule.

The ENN classifier is composed of an input layer of H neurons (where H is the dimension of input space), two hidden layers, and an output layer. The first input layer is composed of I units, whose weights vectors are prototypes $\mathbf{p}_1, \ldots, \mathbf{p}_I$ in input space. The second hidden layer computes mass functions m_i representing the evidence of each prototype $\mathbf{p}_i$. The third layer combines the I mass functions $m_1, \ldots, m_I$ using Dempster's rule. A short description of the ENN model is given in Section 2.5.1.2 section.

The idea of applying the above model to features extracted by a convolutional neural network (CNN) was first proposed by Tong et al. in [Tong et al., 2021a]. In this approach, the ENN module becomes an "evidential layer", which is plugged into the output of a CNN instead of the usual softmax layer. The feature extraction and evidential modules are trained simultaneously. A similar approach was applied in [Tong et al., 2021b] to semantic segmentation. In the next section, we present an alternative approach based on a radial basis function (RBF) network and weights of evidence.

5.2.1.2 *Radial basis function (RBF) network*

As shown in [Denœux, 2019b], the calculations performed in the softmax layer of a feedforward neural network can be interpreted in terms of a combination of evidence by Dempster's rule. The output class probabilities can be seen as normalized plausibilities according to an underlying belief function. Applying these ideas to an RBF network, it is possible to derive an alternative evidential classifier with properties similar to those of the ENN model recalled in Section 2.5.1.2. Consider an RBF network with I prototype (hidden) units. The activation of hidden unit i is

$$s_i = \exp(-\gamma_i d_i^2), \tag{5.1}$$

where, as before, $d_i = \|\mathbf{x} - \mathbf{p}_i\|$ is the Euclidean distance between input vector $\mathbf{x}$ and prototype $\mathbf{p}_i$, and $\gamma_i > 0$ is a scale parameter. For the application considered in this chapter, we only need to consider the case of binary classification with $C = 2$ and $\Omega = \{\omega_1, \omega_2\}$. (The case where $C > 2$ is also analyzed in [Denœux, 2019b]). Let v_i be the weight of the connection between hidden unit i and the output unit, and let $w_i = s_i v_i$ be the product of the output of unit i and weight v_i. The quantities w_i can be interpreted as weights of evidence for class ω_1 or ω_2, depending on the sign of v_i:

- If $v_i \geqslant 0$, w_i a weight of evidence for class ω_1;
- If $v_i < 0$, $-w_i$ is a weight of evidence for class ω_2.

To each prototype, i can, thus, be associated with the following simple mass function:

$$m_i = \{\omega_1\}^{w_i^+} \oplus \{\omega_2\}^{w_i^-},$$

where $w_i^+ = \max(0, w_i)$ and $w_i^- = -\min(0, w_i)$ denote, respectively, the positive and negative parts of w_i. Combining the evidence of all prototypes in favor of ω_1 or ω_2 by Dempster's rule, we get the mass function

$$m = \bigoplus_{i=1}^{I} m_i = \{\omega_1\}^{w^+} \oplus \{\omega_2\}^{w^-}, \tag{5.2}$$

with $w^+ = \sum_{i=1}^{I} w_i^+$ and $w^- = \sum_{i=1}^{I} w_i^-$. In [Denœux, 2019b], the normalized plausibility of ω_1 corresponding to mass function m was shown to have the following expression:

$$p(\omega_1) = \frac{Pl(\{\omega_1\})}{Pl(\{\omega_1\}) + Pl(\{\omega_2\})} = \frac{1}{1 + \exp(-\sum_{i=1}^{I} v_i s_i)}, \tag{5.3}$$

i.e., it is the output of a unit with a logistic activation function. When training an RBF network with a logistic output unit, we thus actually combine evidence from each of the prototypes, but the combined mass function remains latent. In [Denœux, 2019b], mass function m defined by (5.2) was shown to have the following expression:

$$m(\{\omega_1\}) = \frac{[1-\exp(-w^+)]\exp(-w^-)}{1-\kappa} \tag{5.4a}$$

$$m(\{\omega_2\}) = \frac{[1-\exp(-w^-)]\exp(-w^+)}{1-\kappa} \tag{5.4b}$$

$$m(\Omega) = \frac{\exp(-w^+-w^-)}{1-\kappa} = \frac{\exp(-\sum_{i=1}^{I}|w_i|)}{1-\kappa}, \tag{5.4c}$$

where

$$\kappa = [1-\exp(-w^+)][1-\exp(-w^-)] \tag{5.4d}$$

is the degree of conflict between mass functions $\{\omega_1\}^{w^+}$ and $\{\omega_2\}^{w^-}$.

In this approach, we thus need to train a standard RBF network with I prototype units and one output unit with a logistic activation function by minimizing a loss function. Here we define it as the regularized cross-entropy loss

$$L_{CE}(\theta) = -\sum_{n=1}^{N}(y_n \log p_n + (1-y_n)\log(1-p_n)) + \lambda\sum_{i=1}^{I} w_i^2, \tag{5.5}$$

where p_n is the normalized plausibility of class ω_1 computed from (5.3) for instance n, y_n is class label of instance n ($y_n = 1$ if the true class of instance n is ω_1, and $y_n = 0$ otherwise), and λ is a hyperparameter. We note that increasing λ has the effect of decreasing the weights of evidence from prototypes and, thus, obtaining less informative mass functions.

5.2.1.3 *Comparison between ENN and RBF network*

To compare the uncertainty quantification performance of the two classifiers (ENN and RBF network), we consider the two-class dataset shown in Figure 5.2 and show the test error rate and the mean uncertainty of the two classifiers. Furthermore, we also plot the contours of the mass assigned to given classes and the uncertainty. The two classes are randomly distributed around half circles with Gaussian noise and are separated by a nonlinear boundary. A learning set of size $N = 300$ and a test set of size 1000 were generated from the same distribution.

An ENN and an RBF network were initialized with $I = 6$ prototypes generated by the k-means algorithm and were trained on the learning data. Figures 5.3a and 5.3b show, respectively, the test error rate and the mean uncertainty (defined as the average mass assigned to the frame Ω), as functions of hyperparameter λ in (2.21) and (5.5), for 10 different runs of both algorithms with different initializations. As expected, uncertainty increases with λ for both models, but the ENN model appears to be less

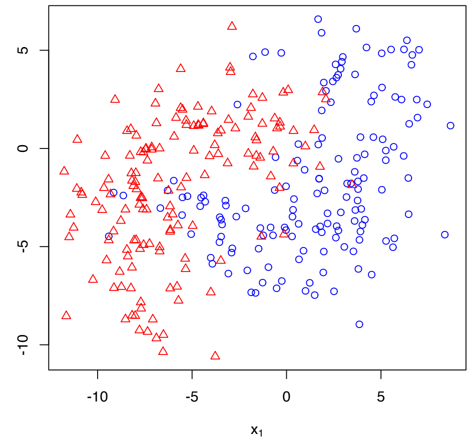


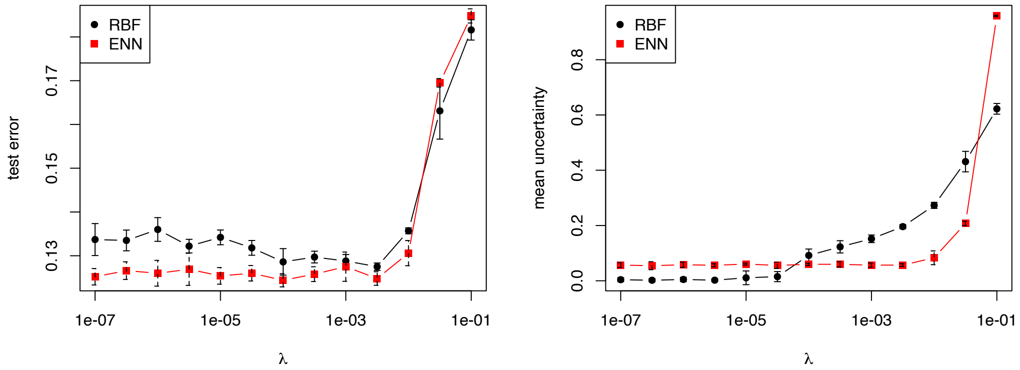
Figure 5.2: Simulated data.

sensitive to λ as compared to the RBF model. Both models achieve similar minimum error rates for λ around 10^{-3}, and have similar mean uncertainties for $\lambda = 10^{-4}$.

As shown in [Denœux, 2000], the robustness of the ENN model arises from the fact that, when the input x is far from all prototypes, the output mass function m is close to the vacuous mass function. This property, in particular, makes the network capable of detecting observations generated from a distribution that is not represented in the learning set. From (5.4c), we can expect the RBF network model to have a similar property: if x is far from all prototypes, all weights of evidence w_i will be small and the mass $m(\Omega)$ will be close to unity. To compare the mass functions computed by the two models, not only in regions of high density where training data are present, but also in regions of low density, we introduced a third class in the test set, as shown in Figure 5.4. Figure 5.5 shows scatter plots of masses on each of the focal sets computed for the two models trained with $\lambda = 10^{-3}$ and applied to an extended dataset composed of the learning data and the third class. We can see that the mass functions are quite similar. Contour plots shown in Figure 5.4 confirm this similarity.

5.2.2 *Uncertainty quantification in evidential segmentation*

We propose an evidential segmentation model to delineate lymphomas, shown in Figure 5.6. It is composed of an encoder-decoder feature extraction module, and an evidential layer based on one of the two models, ENN and RBF network. The input is the concatenated PET-CT image volume provided as a tensor of size $2 \times 256 \times 256 \times 128$, where 2 corresponds to the number of modality channels, and $256 \times 256 \times 128$ is the

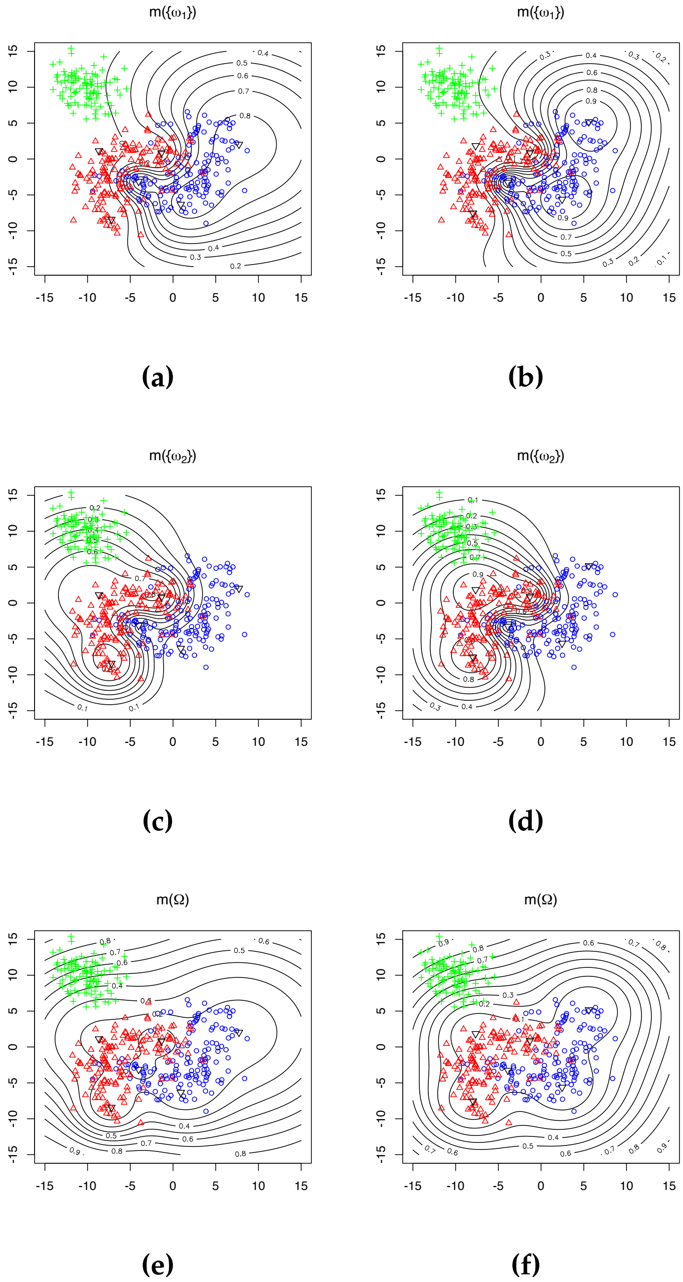


(a) **(b)**

Figure 5.3: Test error rates (a) and mean uncertainty (b) for the ENN and RBF models, as functions of regularization parameter λ.

size of each input volume. The PET-CT image volumes are first fed into the feature extraction module, which outputs high-level features in the form of a tensor of size $256 \times 256 \times 128 \times H$, where H is the number of features computed at each voxel. This tensor is then fed into the evidential layer, which outputs mass functions representing evidence about the class or the ignorance (uncertainty) of each voxel, resulting in a tensor of size $256 \times 256 \times 128 \times (C+1)$, where $C+1$ is the number of masses (one for each class and one for the frame of discernment Ω). The whole network is trained end-to-end by minimizing a regularized Dice loss. The different components of this model are described in greater detail below.

FEATURE EXTRACTION MODULE The feature extraction module is based on a UNet with residual encoder and decoder layers (see Section 1.3.1), as shown in Figure 1.8. Each down-sampling layer (marked in blue) is composed of convolution, normalization, dropout and activation blocks. Each up-sampling layer (marked in green) is composed of transpose convolution, normalization, dropout and activation blocks. The last layer (marked in yellow) is the bottom connection which does not down or up-sample the data. In the experiments reported in Section 5.3, the channels (number of filters) were set as $(8, 16, 32, 64, 128)$ with kernel size equal to 5 and convolutional strides equal to $(2, 2, 2, 2)$. The spatial dimension, input channel and output channel of the module were set, respectively, as 3, 2, and the number H of extracted features. (Experiments with several values of H are reported in Section 5.3.2). The dropout rate was set as 0 and no padding operation was applied. Instance normalization [Ulyanov et al., 2017] was used to perform intensity normalization across the width, height and depth of a single feature map of a single example.

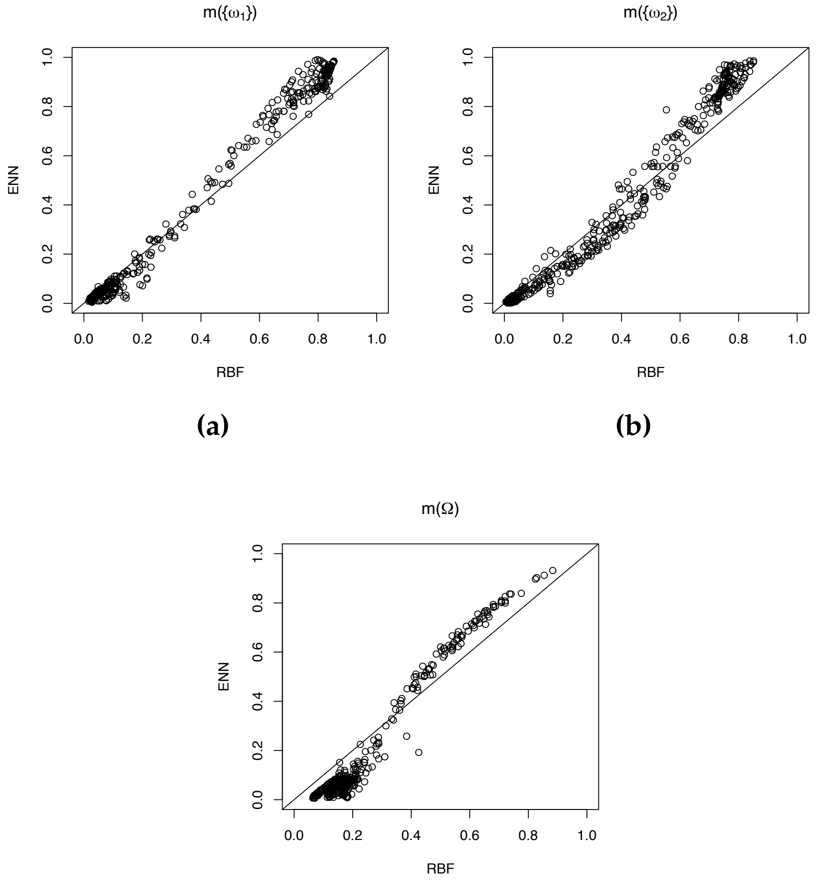


(a) (b)

(c) (d)

(e) (f)

Figure 5.4: Contours of the mass assigned to $\{\omega_1\}$, $\{\omega_2\}$ and Ω by the RBF (left column) and ENN (right column) models. The training data are displayed in blue and red, and the third class (absent from the training data) is shown in green. The training was done with $\lambda = 0.001$ for the two models.



(a)

(b)

(c)

Figure 5.5: Masses computed by the RBF network (horizontal axis) versus the ENN model (vertical axis) for the extended dataset.

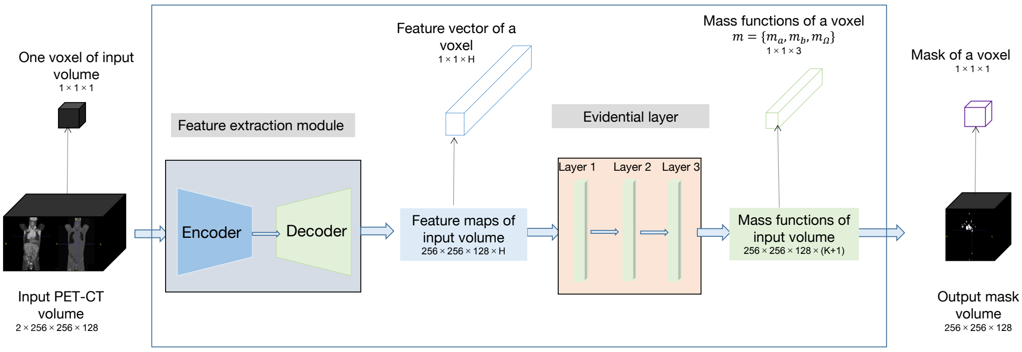


Figure 5.6: Global lymphoma segmentation model.

The Parametric Rectified Linear Unit (PReLU) function [He et al., 2015], which generalizes the traditional rectified unit with a slope for negative values, was used as the activation function. For each input voxel, the feature extraction module outputs a $1 \times H$ feature vector, which is fed into the evidential layer.

EVIDENTIAL LAYER A probabilistic network with a softmax output layer may assign voxels a high probability of belonging to one class while the segmentation uncertainty is actually high because, e.g., the feature vector describing that voxel is far away from feature vectors presented during training. Here, we propose to plug in one of the evidential classifiers described in Sections 5.2.1.1 and 5.2.1.2 at the output of the feature extraction module. The ENN or RBF classifier then takes as inputs the high-level feature vectors computed by the UNet and computes, for each voxel n, a mass function m_n on the frame $\Omega = \{\omega_1, \omega_2\}$, where ω_1 and ω_2 denote, respectively, the background and the lymphoma class, and Ω denotes the uncertainty (see Section 5.3.4 for more details about uncertainty quantification). We will use the names "ENN-UNet" and "RBF-UNet" to designate the two variants of the model.

LOSS FUNCTION The whole network is trained end-to-end by minimizing a regularized Dice loss (see Section 1.2.3). We use the Dice loss instead of the original cross-entropy loss in UNet because the quality of the segmentation is finally assessed by the Dice coefficient. The Dice loss is defined as

$$\mathrm{loss}_D = 1 - \frac{2\sum_{n=1}^{N} S_n G_n}{\sum_{n=1}^{N} S_n + \sum_{n=1}^{N} G_n}, \tag{5.6}$$

where N is the number of voxels in the image volume, S_n is the output pignistic probability of the tumor class (i.e., $m_n(\{\omega_2\}) + m_n(\Omega)/2$) for voxel n, and G_n is ground truth for voxel n, defined as $G_n = 1$ if voxel n corresponds to a tumor, and $G_n = 0$ otherwise. The regularized loss function is

$$\mathrm{loss} = \mathrm{loss}_D + \lambda R, \tag{5.7}$$

where λ is the regularization coefficient and R is a regularizer defined either as $R = \sum_i \alpha_i$ (2.21) if the ENN classifier is used in the ES module or as $R = \sum_i v_i^2$ (5.5) if the RBF classifier is used. The regularization term allows us to decrease the influence of unimportant prototypes and avoid overfitting.

5.3 EXPERIMENTS AND RESULTS

The model introduced in Section 5.2.2 was applied to a set of PET-CT data recorded on patients with lymphomas[1]. The experimental settings are first described in Section 5.3.1. A sensitivity analysis with respect to the main hyperparameters is first reported in Section 5.3.2. We then compare the segmentation accuracy and reliability of our models with those of state-of-the-art methods in Sections 5.3.3 and 5.3.4, respectively.

5.3.1 *Experiment settings*

DATASET The dataset considered in this chapter contains 3D images from 173 patients who were diagnosed with large B-cell lymphomas and underwent PET-CT examination. (The study was approved as a retrospective study by the Henri Becquerel Center Institutional Review Board, Rouen, France). The lymphomas in mask images were delineated manually by experts and considered as ground truth. Figure 5.7 shows examples of PET and CT image slices for one patient with lymphomas. As can be seen, lymphomas in PET images usually correspond to the brightest pixels, but organs such as the brain and bladder are also located in bright pixel areas, which may result in segmentation errors. Moreover, lymphoma boundaries are blurred, which makes it hard to delineate lymphomas precisely. All PET/CT data were stored in the DICOM (Digital Imaging and Communication in Medicine) format. The size and spatial resolution of PET and CT images and the corresponding mask images vary due to the use of different imaging machines and operations. For CT images, the size varies from $267 \times 512 \times 512$ to $478 \times 512 \times 512$. For PET images, the size varies from $276 \times 144 \times 144$ to $407 \times 256 \times 256$.

PRE-PROCESSING Several pre-processing methods were used to process the PET/CT data. At first, the data in DICOM format were transferred into the NIFTI (Neuroimaging Informatics Technology Initiative) format for further processing. Second, the PET, CT and mask images were normalized: (1) for PET images, we applied a random intensity shift and scale of each channel with the shift value of 0 and scale value of 0.1; (2) for CT images, the shift and scale values were set to 1000 and 1/2000; (3) for mask images, the intensity value was normalized into the $[0, 1]$ interval by replacing the outside value by 1. Third, PET and CT images were resized to $256 \times 256 \times 128$ by linear interpolation, and mask images were resized to $256 \times 256 \times 128$ by nearest neighbor interpolation. Lastly, the registration of CT and PET images was performed by B-spline interpolation. All the prepossessing methods can be found in the SimpleITK [Lowekamp et al., 2013][Yaniv et al., 2018] toolkit. During training, PET and CT images were

[1] The code is available at https://github.com/iWeisskohl.

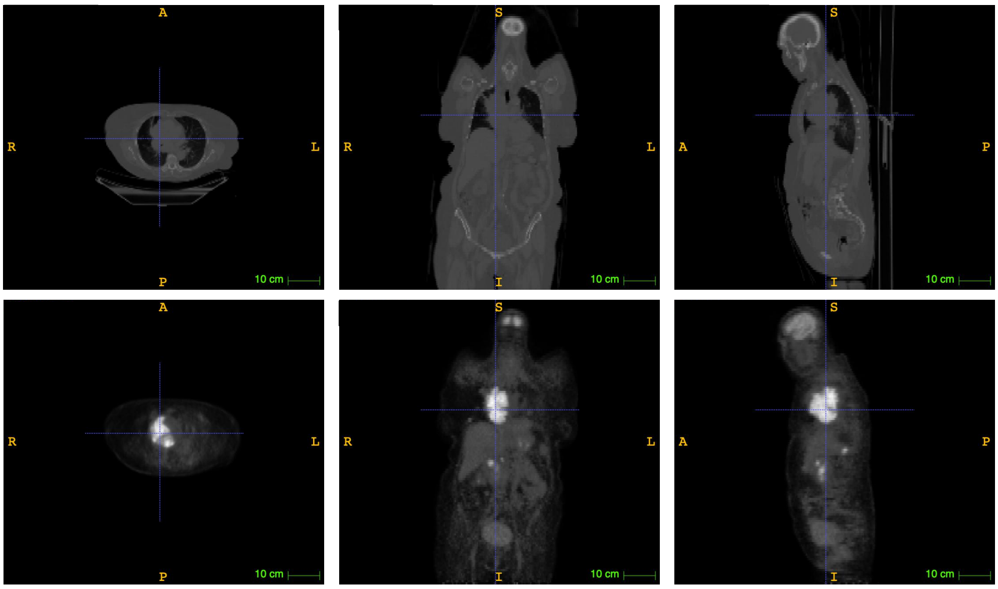


Figure 5.7: Example of a patient with lymphomas. The first and second rows show, respectively, CT and PET slices for one patient in axial, sagittal and coronal views.

concatenated as a two-channel input. We randomly selected 80% of the data for training, 10% for validation, and 10% for testing. This partition was fixed and used in all the experiments reported below.

PARAMETER INITIALIZATION For the evidential layer module, we considered two variants based on the ENN classifier recalled in Section 5.2.1.1 on the one hand, and an RBF network as described in Section 5.2.1.2 on the other hand. Both approaches are based on prototypes in the space of features extracted by the UNet module. Two ways can be used to initialize the prototypes: k-means initialization and random initialization. When using ENN or RBF classifiers as stand-alone classifiers, prototypes are usually initialized by a clustering algorithm such as the k-means. Here, this approach is not so easy, because the whole network is trained in an end-to-end way, and the features are constructed during the training process. However, k-means initialization can still be performed by a four-step process:

1. A standard UNet architecture (with a softmax output layer) is trained end-to-end;

2. The k-means algorithm is run in the space of features extracted by the trained UNet;

3. The evidential layer is trained alone, starting from the initial prototypes computed by the k-means;

4. The whole model (feature extraction module and evidential layer) is fine-tuned by end-to-end learning with a small learning step.

As an alternative method, we also considered training the feature extraction module and the evidential layer simultaneously, in which case the prototypes were initialized randomly from a normal distribution with zero mean and identity covariance matrix. For the ENN module, the initial values of parameters α_i and γ_i were set, respectively, at 0.5 and 0.01, and membership degrees u_{ik} were initialized randomly by drawing uniform random numbers and normalizing. For the RBF module, the initial value of the scale parameter γ_i of RBF was set to 0.01, and the weight v_i was drawn randomly from a standard normal distribution.

LEARNING ALGORITHM Each model was trained on the learning set with 100 epochs using the Adam optimization algorithm. The initial learning rate was set to 10^{-3}. An adjusted learning rate schedule was applied by reducing the learning rate when the training loss did not decrease in 10 epochs. The model with the best performance on the validation set was saved as the final model for testing. All methods were implemented in Python with the PyTorch-based medical image framework MONAI, and were trained and tested on a desktop with a 2.20GHz Intel(R) Xeon(R) CPU E5-2698 v4 and a Tesla V100-SXM2 graphics card with 32 GB GPU memory.

EVALUATION CRITERIA Dice score, Sensitivity and Precision were used to assess the quality of medical image segmentation algorithms. These criteria are defined as follows in (1.1), (1.2) and (1.3), respectively. In addition to the quality of the segmentation, we also wish to evaluate the reliability of output probabilities or belief functions. For that purpose, we used an additional evaluation criterion, the Expected Calibration Error (ECE) [Guo et al., 2017]. The output pignistic probabilities from the evidential layer are first discretized into R equally spaced bins B_r, $r = 1, \ldots, R$ (we used $R = 10$). The accuracy of bin B_r is defined as

$$\mathrm{acc}(B_r) = \frac{1}{\mid B_r \mid} \sum_{i \in B_r} \mathbf{1}(P_i = G_i), \tag{5.8}$$

where P_i and G_i are, respectively, the predicted and true class labels for sample i. The average confidence of bin B_r is defined as

$$\mathrm{conf}(B_r) = \frac{1}{\mid B_r \mid} \sum_{i \in B_r} S_i, \tag{5.9}$$

where S_i is the confidence for sample i. The ECE is the weighted average of the difference in accuracy and confidence of the bins:

$$\mathsf{ECE} = \sum_{r=1}^{R} \frac{\mid B_r \mid}{N} \mid \mathsf{acc}(B_r) - \mathsf{conf}(B_r) \mid, \tag{5.10}$$

where N is the total number of elements in all bins, and $\mid B_r \mid$ is the number of elements in bin B_r. A model is perfectly calibrated when $\mathsf{acc}(B_r) = \mathsf{conf}(B_r)$ for all $r \in \{1, \ldots, R\}$. Through the bin-size weighting in the ECE metric, the highly confident and accurate background voxels significantly affect the results. Because our dataset has an imbalanced foreground and background proportions, we only considered voxels belonging to the tumor to calculate the ECE, similar to [Jungo et al., 2020][Rousseau et al., 2021]. For each patient in the test set, we defined a bounding box covering the lymphoma region and calculated the ECE in this bounding box. We are interested in the patient-level ECE and thus reported the mean patient ECE instead of the voxel-level ECE (i.e., considering all voxels in the test set to calculate the ECE).

5.3.2 *Sensitivity analysis*

We analyzed the sensitivity of the results to the main design hyperparameters, which are: the number H of extracted features, the number I of prototypes and the regulation coefficient λ (5.7). The influence of the initialization method was also studied. In all the experiments reported in this section as well as in Section 5.3.3, learning in each of the configurations was repeated five times with different random initial conditions.

INFLUENCE OF THE NUMBER OF FEATURES Table 5.1 shows the means and standard deviations (over five runs) of the three performance indices for ENN-UNet and RBF-UNet with different numbers of features ($H \in \{2, 5, 8\}$). Here the features come from the feature extraction module. The number of prototypes and the regularization coefficient were set, respectively, to $I = 10$ and $\lambda = 0$. The prototypes were initialized randomly. ENN-UNet achieves the highest Dice score and sensitivity with $H = 2$ features, but the highest precision with $H = 8$. However, the differences are small and concern only the third decimal point. Similarly, RBF-UNet had the best values of the Dice score and precision for $H = 5$ features, but again the differences are small. Overall, it seems that only two features are sufficient to discriminate between tumor and background voxels.

INFLUENCE OF THE REGULARIZATION COEFFICIENT In the previous experiment, the networks were trained without regularization. Tables 5.2 and 5.3 show the performances of ENN-UNet and RBF-UNet for different values of λ, with $I = 10$ randomly initialized prototypes and, respectively, $H = 2$ and $H = 8$ inputs. With both settings, ENN-UNet does not benefit from regularization (the best results are obtained with $\lambda = 0$). In contrast,

Table 5.1: Means and standard deviations (over five runs) of the performance measures for different input dimensions H, with $I = 10$ randomly initialized prototypes and $\lambda = 0$. The best values are shown in bold.

Model	H	Dice score		Sensitivity		Precision	
		Mean	SD	Mean	SD	Mean	SD
ENN-UNet	2	0.833	0.009	0.819	0.019	0.872	0.018
	5	0.831	0.012	0.817	0.016	0.870	0.011
	8	0.829	0.006	0.816	0.010	0.877	0.019
RBF-UNet	2	0.824	0.009	0.832	0.008	0.845	0.016
	5	0.825	0.006	0.817	0.016	0.862	0.010
	8	0.821	0.011	0.813	0.010	0.862	0.022

Table 5.2: Means and standard deviations (over five runs) of the performance measures for different values of the regularization coefficient λ, with $I = 10$ randomly initialized prototypes and $H = 2$ features. The best values are shown in bold.

Model	λ	Dice score		Sensitivity		Precision	
		Mean	SD	Mean	SD	Mean	SD
ENN-UNet	0	0.833	0.009	0.819	0.019	0.872	0.018
	1e-4	0.822	0.007	0.818	0.026	0.839	0.035
	1e-2	0.823	0.004	0.817	0.023	0.856	0.023
RBF-UNet	0	0.824	0.009	0.832	0.008	0.845	0.016
	1e-4	0.825	0.011	0.811	0.022	0.869	0.020
	1e-2	0.829	0.010	0.818	0.022	0.867	0.016

RBF-UNet is more sensitive to regularization and achieves the highest Dice score with $\lambda = 0.01$. This finding confirms the remark already made in Section 5.2.1.3, where it was observed that an ENN classifier seems to be less sensitive to regularization than an RBF classifier (see Figure 5.3a).

influence of the number of prototypes The number I of prototypes is another hyperparameter that may impact segmentation performance. Table 5.4 shows the performances of ENN-UNet and RBF-UNet with 10 and 20 randomly initialized prototypes, the other hyperparameters being fixed at $H = 2$ and $\lambda = 0$. Increasing the number of prototypes beyond 10 does not seem to improve the performance of ENN-UNet, while it does slightly improve the performance of RBF-UNet in terms of Dice score and precision, at the expense of an increased computing time.

influence of the prototype initialization method Finally, we compared the two initialization methods mentioned in Section 5.3.1. For k-means initialization, in the first step, a UNet model was trained with the following settings: kernel size=5, channels $=(8, 16, 32, 64, 128)$ and strides$=(2, 2, 2, 2)$. The spatial dimension, input, and output channel were

Table 5.3: Means and standard deviations (over five runs) of the performance measures for different values of the regularization coefficient λ, with $I = 10$ randomly initialized prototypes and $H = 8$ features. The best values are shown in bold.

Model	λ	Dice score		Sensitivity		Precision	
		Mean	SD	Mean	SD	Mean	SD
ENN-UNet	0	0.829	0.006	0.811	0.010	0.877	0.019
	1e-4	0.827	0.008	0.809	0.019	0.873	0.024
	1e-2	0.822	0.009	0.807	0.021	0.867	0.011
RBF-UNet	0	0.821	0.010	0.813	0.010	0.862	0.022
	1e-4	0.827	0.004	0.830	0.005	0.852	0.012
	1e-2	0.832	0.006	0.825	0.022	0.867	0.020

Table 5.4: Means and standard deviations (over five runs) of the performance measures for different numbers I of randomly initialized prototypes, with $H = 2$ features and $\lambda = 0$. The best values are shown in bold.

Model	I	Dice score		Sensitivity		Precision	
		Mean	SD	Mean	SD	Mean	SD
ENN-UNet	10	0.833	0.009	0.819	0.019	0.872	0.018
	20	0.823	0.007	0.804	0.006	0.864	0.012
RBF-UNet	10	0.824	0.009	0.832	0.008	0.845	0.016
	20	0.830	0.007	0.810	0.012	0.867	0.010

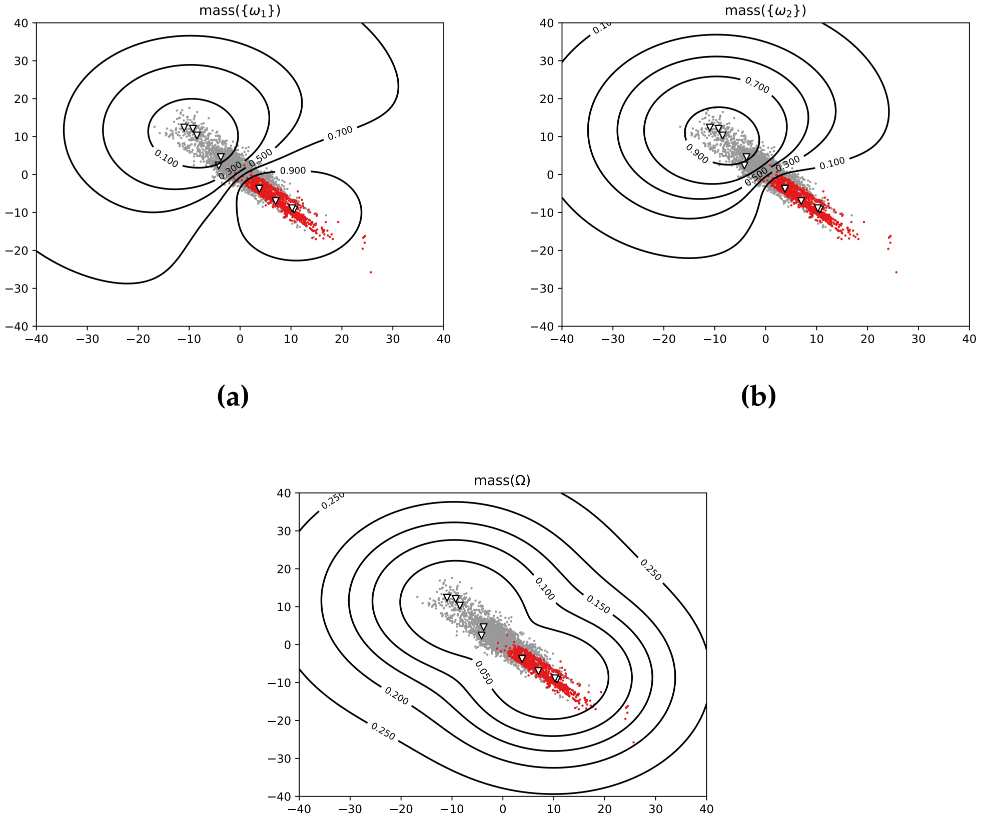
Table 5.5: Means and standard deviations (over five runs) of the performance measures for different initialization methods, with $I = 10$ prototypes, $H = 2$ features and $\lambda = 0$. The best values are shown in bold.

Model	Initialization	Dice score		Sensitivity		Precision	
		Mean	SD	Mean	SD	Mean	SD
ENN-UNet	Random	0.833	0.009	0.819	0.019	0.872	0.018
	k-means	0.846	0.002	0.830	0.004	0.879	0.008
RBF-UNet	Random	0.824	0.009	0.832	0.008	0.845	0.016
	k-means	0.839	0.003	0.824	0.001	0.879	0.008

set, respectively to, 3, 2, and 2. This pre-trained UNet was used to extract $H = 2$ features, and 10 prototypes were obtained by running the k-means algorithm in the space of extracted features. These prototypes were fed into ENN or RBF layers, which were trained separately, with fixed features. For this step, the learning rate was set to 10^{-2}. Finally, the whole model was fine-tuned end-to-end, with a smaller learning rate equal to 10^{-4}. Table 5.5 shows the performances of ENN-UNet and RBF-UNet with random and k-means initialization. Both ENN-UNet and RBF-UNet achieve a higher Dice score when using the k-means initialization method, and the variability of the results is also reduced with this method.

Not only does the k-means initialization method slightly improve the performances of ENN-UNet and RBF-UNet quantitatively, but it also tends to position the prototypes in regions of high data density. As a result, a high output mass $m(\Omega)$ signals that the input data is atypical. In that sense, the output mass function is more interpretable. This point is illustrated by Figures 5.8 and 5.9, which show the contours, in the two-dimensional feature space, of the masses assigned to the background, the tumor class and the frame of discernment when using k-means initialization (with $\lambda = 10^{-2}$ and $I = 10$) with, respectively, ENN-UNet and RBF-UNet. For both models, the prototypes are well distributed over the two classes, and the mass on Ω decreases with the distance to the data, as expected. In contrast, when using random initialization (as shown in Figure 5.10 for the ENN-UNet model – results are similar to the RBF-UNet model), the prototypes are located in the background region, and the mass $m(\Omega)$ does not have a clear meaning (although the decision boundary still ensures good discrimination between the two classes).

From this sensitivity analysis, we can conclude that the performances of both ENN-UNet and RBF-UNet are quite robust to the values of the hyperparameters and that the two models achieve comparable performances. The k-means initialization method seems to yield better results, both quantitatively and qualitatively. The next section is devoted to a comparison with alternative models.

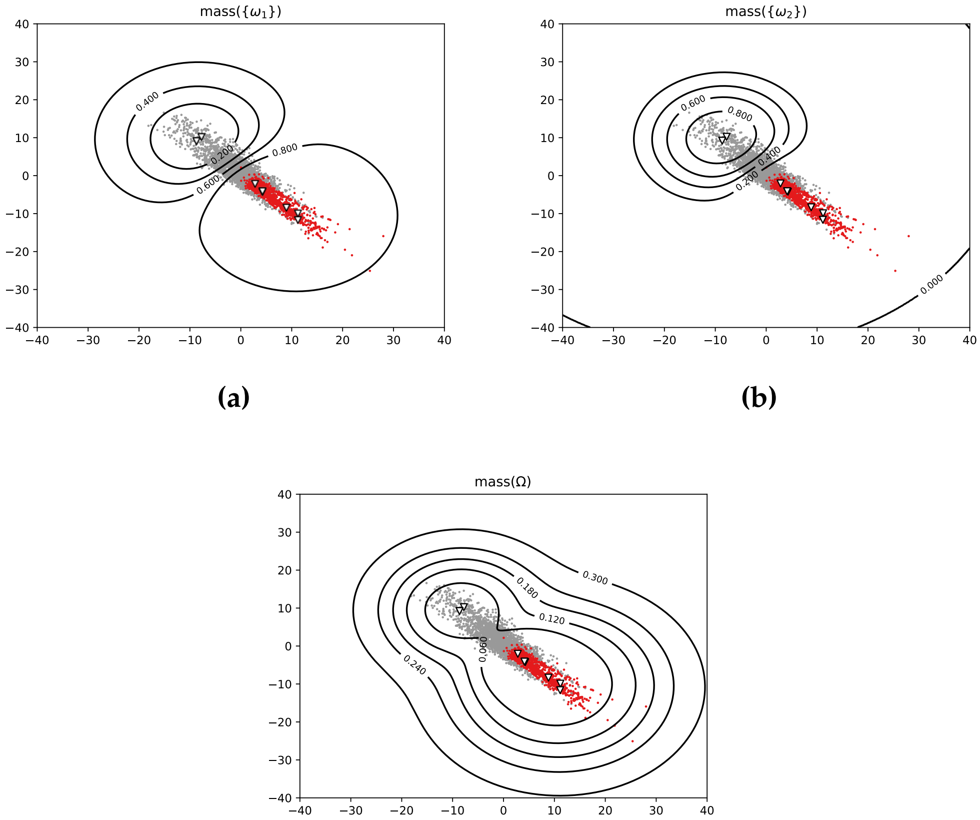


(a)

(b)

(c)

Figure 5.8: Contours in feature space of the masses assigned to the background (a), the tumor class (b), and the frame of discernment (c) by the ENN-UNet model initialized by k-means. Training was done with $\lambda = 10^{-2}$, $H = 2$ and $I = 10$. Sampled feature vectors from the tumor and background classes are marked in gray and red, respectively.

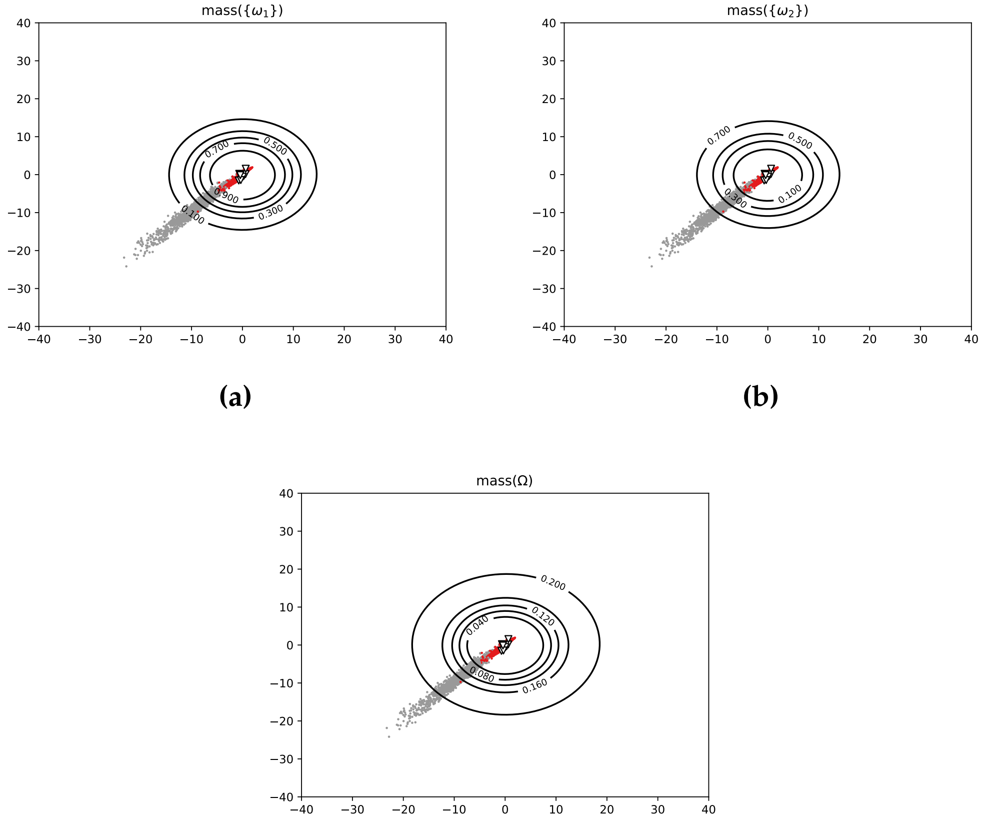


(a)

(b)

(c)

Figure 5.9: Contours in feature space of the masses assigned to the background (a), the tumor class (b) and the frame of discernment (c) by the RBF-UNet model initialized by k-means. Training was done with $\lambda = 10^{-2}$, $H = 2$ and $I = 10$. Sampled feature vectors from the tumor and background classes are marked in gray and red, respectively.



(a)

(b)

(c)

Figure 5.10: Contours in feature space of the masses assigned to the background (a), the tumor class (b), and the frame of discernment (c) by the ENN-UNet model initialized randomly. Training was done with $\lambda = 10^{-2}$, $H = 2$ and $I = 10$. Sampled feature vectors from the tumor and background classes are marked in gray and red, respectively.

5.3.3 *Comparative analysis: segmentation accuracy*

In this section, we compare the performances of the ENN-UNet and RBF-UNet models with those of the baseline model, UNet [Ronneberger et al., 2015], as well as three state-of-the-art models: VNet [Milletari et al., 2016], SegResNet [Myronenko, 2018] and nnUNet [Isensee et al., 2018]. VNet is a variant of UNet that introduces short residual connections at each stage. Compared with UNet, SegResNet contains an additional variational autoencoder branch. nnUNet (see Section 1.3.3) is the first segmentation model designed as a segmentation pipeline for any given dataset. For all compared methods, the same learning set and pre-processing steps were used. All the compared methods were trained with the Dice loss function (5.6). Details about the optimization algorithm were given in Section 5.3.1. All methods were implemented based on the MONAI framework[2] and can be called directly. For UNet, the kernel size was set as 5 and the channels were set to $(8, 16, 32, 64, 128)$ with strides=$(2, 2, 2, 2)$. For nnUNet, the kernel size was set as $(3, (1, 1, 3), 3, 3)$ and the upsample kernel size was set as $(2, 2, 1)$ with strides $((1, 1, 1), 2, 2, 1)$. For SegResNet [Myronenko, 2018] and VNet [Milletari et al., 2016], we used the pre-defined model without changing any parameter. The spatial dimension, input channel and output channel were set, respectively, 3, 2, and 2 for the four compared models. As for other hyperparameters not mentioned here, we used the pre-defined value given in MONAI. As shown by the sensitivity analysis performed in Section 5.3.2, the best results for ENN-UNet and RBF-UNet are achieved with $\lambda = 0$, $I = 10$, $H = 2$ and k-means initialization.

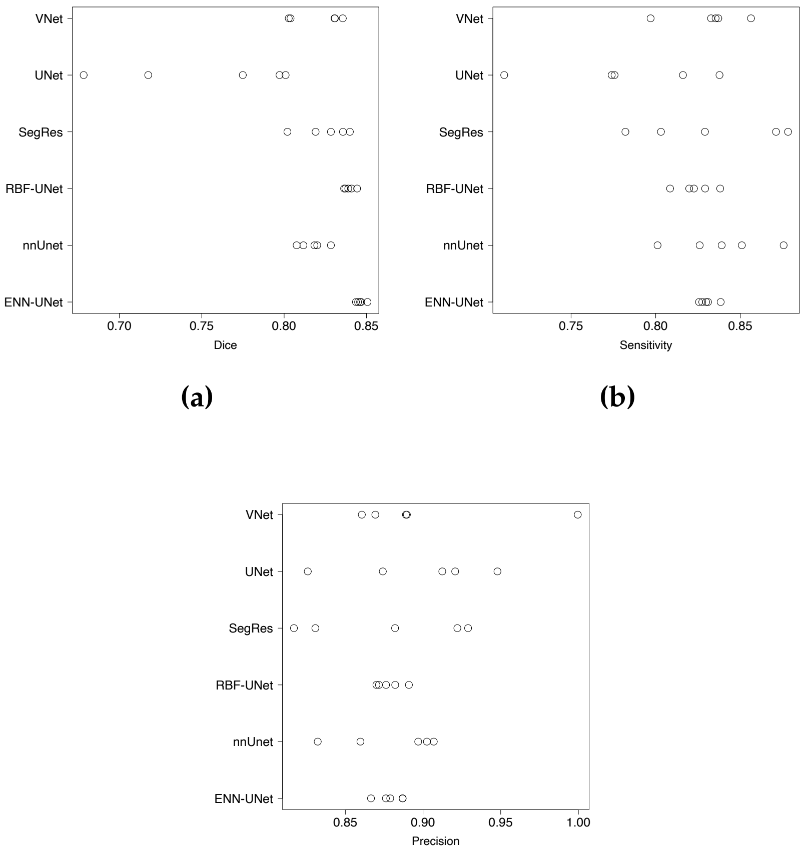
The means and standard deviations of the Dice score, sensitivity, and precision over five runs with random initialization for the six methods are shown in Table 5.6, and the raw values are plotted in Figure 5.11. We can see that ENN-UNet and RBF-UNet achieve, respectively, the highest and the second-highest mean Dice scores. A Kruskal-Wallis test performed on the whole data concludes with a significant difference between the distributions of the Dice score for the six methods (p-value = 0.0001743), while the differences are not significant for sensitivity (p-value = 0.2644) and precision (p-value = 0.9496). Table 5.7 shows the results of the Conover-Iman test of multiple comparisons [Conover et al., 1979][Dinno et al., 2017] with Benjamini-Yekutieli adjustment [Benjamini et al., 2001]. We can see that the differences between the Dice scores obtained by ENN-UNet and RBF-UNet on the one hand, and the four other methods on the other hand are highly significant (p-values $< 10^{-2}$), while the difference between ENN-UNet and RBF-UNet is only weakly significant (p-value = 0.0857).

Figure 5.12 shows two examples of segmentation results obtained by ENN-UNet and UNet, corresponding to large and isolated lymphomas.

[2] More details about how to use those models can be found from MONAI core tutorials https://monai.io/started.html#monaicore.

Table 5.6: Means and standard deviations (over five runs) of the performance measures for ENN-UNet, RBF-UNet and four reference methods. The best result is shown in bold, and the second best is underlined.

Model	Dice score		Sensitivity		Precision	
	Mean	SD	Mean	SD	Mean	SD
UNet [Kerfoot et al., 2018]	0.753	0.054	0.782	0.048	0.896	0.047
nnUNet [Isensee et al., 2018]	0.817	0.008	**0.838**	0.028	0.879	0.032
VNet [Milletari et al., 2016]	0.820	0.016	0.831	0.021	**0.901**	0.056
SegResNet [Myronenko, 2018]	0.825	0.015	0.832	0.042	0.876	0.051
ENN-UNet	**0.846**	0.002	0.830	0.004	0.879	0.008
RBF-UNet	0.839	0.003	0.824	0.001	0.879	0.008



(a)

(b)

(c)

Figure 5.11: Values of the Dice score (a), sensitivity (b) and precision (c) for five runs of the six methods.

Table 5.7: Conover-Iman test of multiple comparisons between the Dice scores obtained by the six models: t-test statistics and p-values. P-values less than 0.01 are printed in bold.

	ENN-UNet	nnUnet	RBF-UNet	SegResNet	UNet
nnUnet	6.759 **0.0000**				
RBF-UNet	2.156 0.0857	-4.602 **0.0004**			
SegResNet	5.349 **0.0001**	-1.410 0.3282	3.193 **0.0088**		
UNet	10.283 **0.0000**	3.524 **0.0043**	8.127 **0.0000**	4.934 **0.0002**	
VNet	6.054 **0.0000**	-0.705 0.8091	3.898 **0.0019**	0.705 0.8669	-4.229 **0.0009**

We can see, in these two examples, that UNet is more conservative (it correctly detects only a subset of the tumor voxels), which may explain why it has relatively high precision. However, the tumor regions predicted by ENN-UNet better overlap the ground-truth tumor region, which is also reflected by the higher Dice score.

5.3.4 *Comparative analysis: segmentation uncertainty*

Besides segmentation accuracy, another important issue concerns the quality of uncertainty quantification. Monte-Carlo dropout (MCD) [Gal et al., 2016] is a state-of-the-art technique for improving the uncertainty quantification capabilities of deep networks. In this section, we compare the ECE (5.10) achieved by UNet (the baseline), SegResNet (the best alternative method found in Section 5.3.3), and our proposals: ENN-UNet, and RBF-UNet, with and without MCD. For the four methods, the dropout rate was set to 0.5 and the sample number was set to 20; we averaged the 20 output probabilities (the pignistic probabilities for the two evidential models) at each voxel as the final output of the model.

The results are reported in Table 5.8. We can see that MCD enhances the segmentation performance (measured by the Dice index) of UNet et SegResNet, and improves the calibration of all methods, except SegResNet. Overall, the smallest average ECE is achieved by RBF-UNet and ENN-UNet with MCD, but the standard deviations are quite large. A Kruskal-Wallis test concludes with a significant difference between the distributions of ECE for the eight methods (p-value = 0.01). The p-values of the Conover-Iman test of multiple comparisons with Benjamini-Yekutieli adjustment reported in Table 5.9 show significant differences between the ECE of RBF-UNet with MCD on the one hand, and those of RBF-UNet without MCD, SegResNet with MCD, and UNet without MCD on the other hand. We

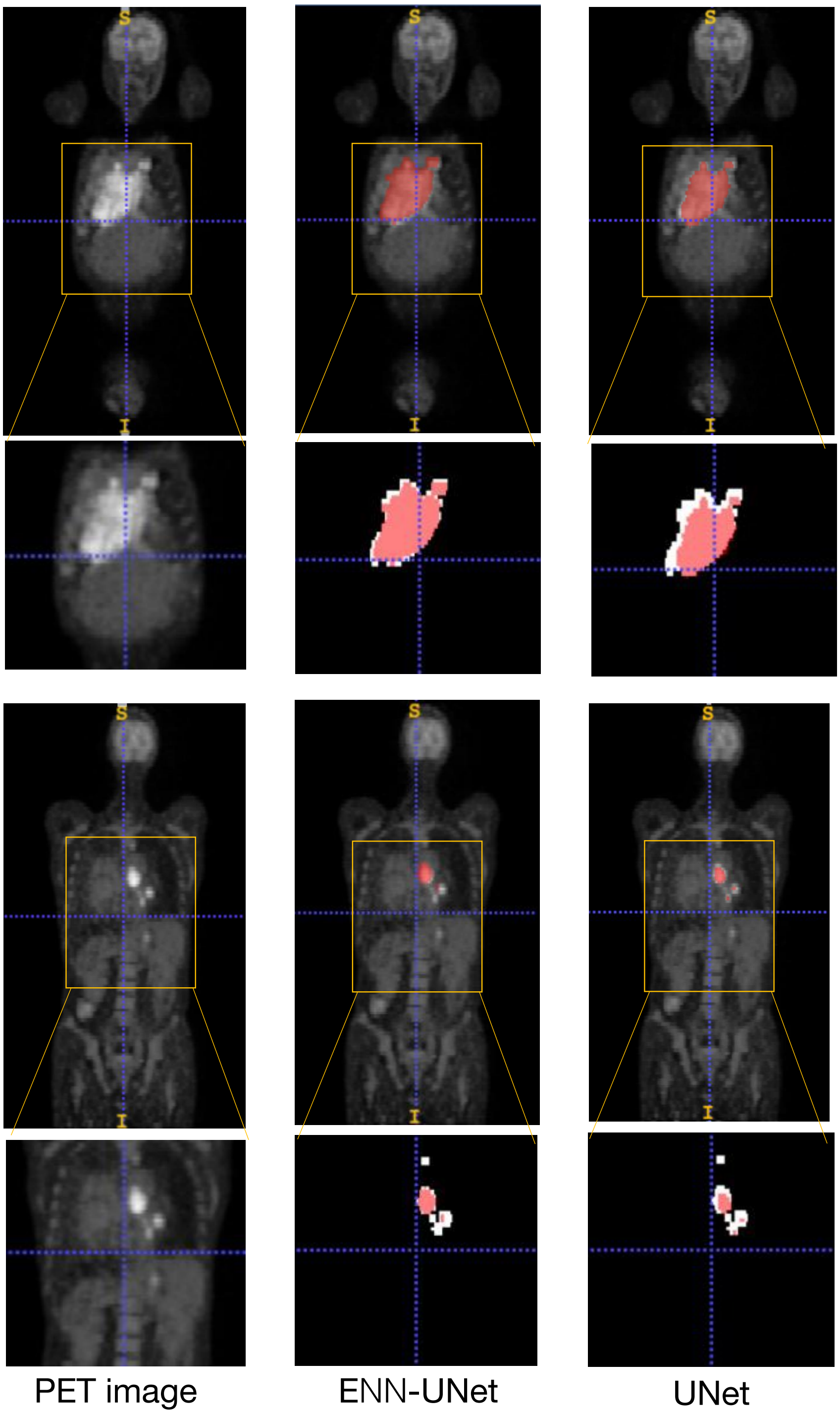


Figure 5.12: Two examples of segmentation results by ENN-UNet and UNet. The first and the second row are, respectively, representative of large and isolated small lymphomas. The three columns correspond, from left to right, to the PET images and the segmentation results obtained by ENN-UNet and UNet. The white and red regions represent, respectively, the ground truth and the segmentation result.

Table 5.8: Means and standard deviations (over five runs) of the Dice score and ECE for UNet, SegResNet, ENN-UNet, and RBF-UNet, with and without MCD. The best results are shown in bold, and the second best is underlined.

Model	Dice score		ECE(%)	
	Mean	SD	Mean	SD
UNet	0.754	0.054	2.22	0.205
SegResNet	0.825	0.015	1.97	0.488
ENN-UNet	**0.846**	0.002	1.99	0.110
RBF-UNet	0.839	0.003	2.12	0.028
UNet with MCD	0.828	0.005	1.93	0.337
SegResNet with MCD	0.844	0.009	2.53	0.973
ENN-UNet with MCD	0.841	0.003	1.53	0.075
RBF-UNet with MCD	0.840	0.003	**1.52**	0.041

Table 5.9: Conover-Iman test of multiple comparisons between the ECE obtained by UNet, SegResNet, ENN and RBF, with and without MCD: t-test statistics and p-values. P-values less than 0.01 are printed in bold.

	ENN	ENN-MC	RBF	RBF-MC	SegRes	SegRes-MC	UNet
ENN-MC	0.926						
	1.0000						
RBF	-1.191	-2.118					
	0.7403	0.2892					
RBF-MC	2.812	1.886	4.004				
	0.1145	0.3419	**0.0095**				
SegRes	0.695	-0.232	1.886	-2.117			
	1.0000	1.0000	0.3761	0.3305			
SegRes-MC	-0.860	-1.787	0.331	-3.673	-1.555		
	1.0000	0.3530	1.0000	**0.0159**	0.4756		
UNet	-1.357	-2.283	-0.165	-4.169	-2.051	-0.496	
	0.6337	0.2677	1.0000	**0.0119**	0.2962	1.0000	
UNet-MC	0.430	-0.496	1.621	-2.382	-0.265	1.290	1.787
	1.0000	1.0000	0.4507	0.2564	1.0000	0.6667	0.3824

also tested the pairwise differences between the ECE values obtained by RBF-UNet and ENN-UNet with MCD on the one hand, and UNet with and without MCD as well as SegResNet with and without MCD, on the other hand, using the Wilcoxon rank sum test. The corresponding p-values are shown in Table 5.10. We find significant differences between the ECE RBF-UNet with MCD and those of the other methods, but only a weakly significant difference between ENN-UNet with MCD and UNet without MCD. In summary, there is some evidence that MCD improves calibration, even for evidential models, and that the best calibration is achieved by the RBF-UNet model, but this evidence is not fully conclusive due to the limited size of the dataset; our findings will have to be confirmed by further experiments with larger datasets.

Table 5.10: P-values for the Wilcoxon rank sum test applied to the comparison of ECE obtained by ENN-UNet and RBF UNet with MCD on the one hand and the four other methods on the other hand (UNet and SegResNet with and without MCD).

	UNet	UNet-MC	SegRes	SegRes-MC
ENN-MC	0.095	0.67	0.69	0.31
RBF-MC	0.0079	0.012	0.055	0.0079

5.4 CONCLUSION

An evidential model for segmenting lymphomas from 3D PET-CT images with uncertainty quantification has been proposed in this chapter. Our model is based on the concatenation of a UNet, which extracts high-level features from the input images, and an evidential segmentation module, which computes output mass functions for each voxel. Two versions of this evidential module, both involving prototypes, have been studied: one is based on the ENN classifier initially proposed as a stand-alone classifier in [Denœux, 2000], while the other one relies on an RBF classifier and the addition of the weight of evidence. The whole model is trained end-to-end by minimizing the Dice loss. Our model has been shown to outperform the baseline UNet model and other state-of-the-art segmentation methods on a dataset of 173 patients with lymphomas. Preliminary results also suggest the outputs of the evidential models (in particular, the one with an RBF layer) are better calibrated and that calibration error can be further decreased by Monte Carlo dropout. These results, however, will have to be confirmed by further experiments with larger datasets. One of the potential problems that may arise is related to the dimensionality of the feature space. In the application considered in this work, good results were obtained with only two extracted features. If some other learning tasks require a much higher feature dimension, we may need a much higher number of prototypes, and learning may be slow. This issue could be addressed by adapting the loss function as proposed, e.g., in [Hryniowski et al., 2019].

Multimodal medical images indicate different information about the presence of disease, with varying reliability of each modality image. In this work, we use two modalities: PET providing information on the lymphoma and CT providing anatomical information, to better guide radiotherapy planning. We fuse images from PET and CT modalities by concatenating them into a single input and quantifying the segmentation uncertainty. This work can be extended in many research directions. For example, we can further evaluate the approach by applying it to other medical image segmentation problems and study the quantification and fusion of modality-level uncertainty. In the next chapter, we will introduce our work on quantifying modality-level uncertainty and combining unreliable sources of information with an application to a large-scale brain tumor dataset.

Chapter 6

Multimodal medical image segmentation with contextual discounting

Contents

6.1 introduction

Single-modality medical images often do not contain enough information to reach an accurate and reliable diagnosis. This is why physicians generally use multiple sources of information, such as multi-MR images for brain tumor segmentation, or PET-CT images for lymphoma segmentation. The effective fusion of multimodal information is of great importance in the medical domain for better disease diagnosis and radiotherapy. Using convolutional neural networks (CNNs), researchers have mainly adopted probabilistic approaches to information fusion, which can be classified into three strategies: image-level fusion, such as input data concatenation [Peiris et al., 2021]; feature-level fusion, such as attention mechanism concatenation [Zhou et al., 2020]; and decision-level fusion such as weighted averaging [Kamnitsas et al., 2017]. However, probabilistic fusion is unable to effectively manage conflict that occurs when different labels are assigned to the same voxel based on different modalities. Also, it is important, when combining multimodal information, to take into account the reliability of the different sources of information.

In this chapter, we introduce a multimodal evidence fusion framework based on contextual discounting [Pichon et al., 2016; Mercier et al., 2012] and deep learning. In the BFT framework, the reliability of a source of information can be taken into account using the discounting operation [Shafer, 1976], which transforms each piece of evidence provided by a source into a weaker, less informative one. To our knowledge, this work is the first attempt to apply evidence theory with contextual discounting to the fusion of deep neural networks. The idea of considering multimodal images as independent inputs and quantifying their reliability is simple and reasonable. However, modeling the reliability of sources is important and challenging. Our method computes mass functions to assign degrees of belief to each class and assign an ignorance degree to the whole set of classes. It thus has one more degree of freedom than a probabilistic model, which allows us to model source uncertainty directly. The contributions of this work are the following: (1) Four BFT-based evidential segmentation modules are used to compute the belief of each voxel belonging to the tumor for four modality MR images; (2) An evidence-discounting mechanism is applied to each of the single-modality MR images to take into account its reliability; (3) A multimodal evidence fusion strategy is then applied to combine the discounted evidence with BFT and achieve more reliable results. End-to-end learning is performed by minimizing a new loss function based on the discounting mechanism, allowing us to increase the segmentation performance and reliability. The framework is evaluated on the BraTs 2021 database of 1251 patients with brain tumors. Quantitative and qualitative results show that our method outperforms the state-of-the-art and implements an effective new idea for merging multiple information within deep neural networks.

We organize this chapter as follows: Section 6.2 introduces the multimodal medical image segmentation model. Section 6.3 reports the numerical experiments by evaluating the performance in segmentation accuracy and reliability. Finally, we conclude this work in Section 6.4.

6.2 PROPOSED APPROACH

Our proposed evidence fusion framework is shown in Figure 6.1. It is composed of (1) four encoder-decoder feature extraction (FE) modules, (2) four evidential segmentation (ES) modules, and (3) a multimodal evidence fusion (MMEF) module. Since this chapter focuses on uncertainty representation and multimodal evidence fusion, the FE module here can be

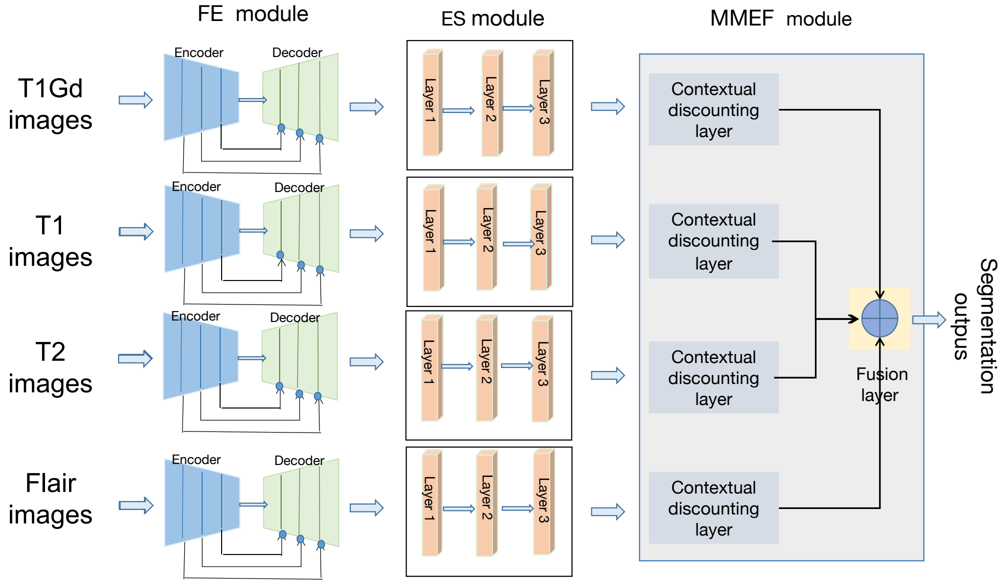


Figure 6.1: Multimodal evidence fusion framework. It is composed of four encoder-decoder feature extraction (FE) modules corresponding to T1Gd, T1, T2 and FLAIR modality inputs; four evidential segmentation (ES) modules corresponding to each of the inputs; and a multimodal evidence fusion (MMEF) module.

any state-of-the-art medical image segmentation models, such as Residual-UNet (see Section 1.3.1), nnUNet (see Section 1.3.3). Details about the evidential segmentation and multimodal evidence fusion modules will be given in Sections 6.2.1 and 6.2.2, respectively. The loss function used for optimizing the framework will be described in Section 6.2.3.

6.2.1 *Evidential segmentation (ES) module*

Based on ENN introduced in 2.5.1.2, we propose a BFT-based evidential segmentation module to quantify the uncertainty about the class of each voxel by a mass function. The basic idea of the evidential segmentation module is to assign a mass to each of the C classes and to the whole set of classes Ω, based on the distance between the input x of each voxel and I prototypes. The input x is the high-level semantic feature vector generated by the feature extraction module. For each prototype p_i, the evidential segmentation module will first calculate mass functions m_i representing the corresponding evidence using the following equations:

$$m_i(\{\omega_c\}) = u_{ic}\alpha_i \exp(-\gamma_i\|x - p_i\|^2), \quad c = 1, ..., C \tag{6.1a}$$

$$m_i(\Omega) = 1 - \alpha_i \exp(-\gamma_i\|x - p_i\|^2), \tag{6.1b}$$

where $\gamma_i > 0$ and $\alpha_i \in [0, 1]$ are two parameters, u_{ic} is the membership degree of prototype i to class ω_c, and $\sum_{c=1}^{C} u_{ic} = 1$. Then the I pieces of evidence are combined by Dempster's rule (see Section 2.2). As shown in Fig 6.1, for each input modality, we have an independent evidential segmentation module plugged into the feature extraction module's output. For each voxel, the evidential segmentation module generates a mass function m to each of the C classes and a mass function $m(\Omega)$ to the whole set of classes Ω. Here $m(\Omega)$ is regarded as the degree of uncertainty when assigning a voxel to the given classes.

6.2.2 *Multimodal evidence fusion (MMEF) module*

In this work, we address the problem of modeling the reliability of single-modality medical images by the discounting operation under the BFT framework. The multiple discounted evidence is then combined by Dempster's rule (see Section 2.2). In this section, we explain the use of contextual discounting in multimodal MR image fusion.

DISCOUNTING SOURCE EVIDENCE WITH RELIABILITY COEFFICIENT We follow Mercier's idea [Mercier et al., 2008] to extend the *discounting* operation into *contextual discounting*. With *contextual discounting*, we can represent richer meta-knowledge regarding the reliability of the single modality image in different contexts, i.e., different types of MR images are conditionally reliable regarding different brain tumor segmentation tasks. We assume that we have evidence regarding the reliability of a source modality S, conditionally on each hypothesis $\omega_c \in \Omega$, i.e., in a context where the quantity x of interest is known to be equal to ω_c. When it is known that the actual value of x is ω_c, $1-\beta_c$ is the plausibility that the source is not reliable in the same context, and thus our knowledge about x taking value in $\{\omega_c\}$ is vacuous, i.e., $m_?$. For each source of information, we thus have a vector $\boldsymbol{\beta} = [\beta_1, ..., \beta_C]$ representing its reliability in different contexts. The contour function of the discounted mass function can be shown [Mercier et al., 2008] to be

$$^{\beta}pl(\{\omega_c\}) = 1 - \beta_c + \beta_c\, pl(\{\omega_c\}), c = 1, ..., C, \tag{6.2}$$

where $^{\beta}pl(\{\omega_c\})$ is the discounted plausibility of class $m(\{\omega_c\})$, $pl(\{\omega_c\}) = m(\{\omega_c\}) + m(\Omega)$. By representing the evidence with the contour function, we can decrease the calculation complexity.

FUSION WITH DISCOUNTED SOURCE EVIDENCE The four independent discounted evidence corresponding to four input modalities must be combined to generate final evidence. Equation (2.8) allows us to compute the contour function in time proportional to the size of Ω, without having to compute the combined mass $m_1 \oplus m_2$ with (2.6). In our case, $^{\beta^1}pl_{T1}$, $^{\beta^2}pl_{T2}$,

${}^{\beta^3}pl_{T1Gd}$ and ${}^{\beta^4}pl_{FLAIR}$ represent the contour functions provided, respectively, by modality T1, T2, T1Gd and FLAIR, with discount rate vectors $1-\beta^1$, $1-\beta^2$, $1-\beta^3$ and $1-\beta^4$. As a consequence of (2.8), the normalized contour function of multiple sources of information is proportional to the product of the contour function of each source information and can be used to simplify the processes of the orthogonal sum of m_1 and m_2 using Dempster's rule. The combined contour function ${}^{\beta}pl$ of each singleton is, thus,

$$ {}^{\beta}pl(\{\omega_c\}) = \frac{\prod_{t=1}^{T} {}^{\beta_c^t}pl_t(\{\omega_c\})}{\sum_{c=1}^{C}\prod_{t=1}^{T} {}^{\beta_c^t}pl_t(\{\omega_c\})}, c = 1, ..., C, \qquad (6.3)$$

where T is the number of information sources, and C is the number of classes on Ω.

6.2.3 *Learning with contextual discounted Dice loss*

To optimize the whole framework, we maximize the overlap region between the output S and the ground truth G by minimizing the following loss function $loss_D$, given as follows:

$$loss_D = 1 - \frac{2\sum_{n=1}^{N}\sum_{c=1}^{C} S_{cn} \times G_{cn}}{\sum_{n=1}^{N}\sum_{c=1}^{C} S_{cn} + G_{cn}}, \qquad (6.4)$$

where C and N are the number of classes and voxels; $G_{cn} = 1$ if voxel n belongs to class c, and $G_{cn} = 0$ otherwise, and S_{cn} represents the corresponding predicted mask based on the discounted source evidence after fusion.

6.3 EXPERIMENTS AND RESULTS

6.3.1 *Experiment settings*

DATASET We used a multimodal brain tumor dataset from the BraTS 2021 challenge [Baid et al., 2021] to evaluate our framework. The original BraTS2021 dataset comprises training, validation and test sets with the corresponding number of cases as 1251, 219 and 570, respectively. Compared to the BraTS2019 dataset we used in Chapter 4, BraTS2021 is a more recent version with more brain tumor cases available. For BraTS2021, the validation and test set and the corresponding ground truth are not available; thus, in this chapter, we train and evaluate our framework with the training set. Following [Peiris et al., 2021], we randomly divided the 1251 scans into 834, 208, and 209 for new training, validation, and test set, respectively. Each case has four modalities: T1, T1Gd, T2, and FLAIR with $240 \times 240 \times 155$ voxels. Figure 6.2 shows four modality MR image slices from one patient. The appearance of a brain tumor varies in different modalities, and the tumor boundaries are blurred, making it hard to delineate different tumors precisely. Annotations of 1251 scans comprise the GD-enhancing tumor (ET-label 4), the peritumoral edema (ED-label 2), and the necrotic and non-enhancing tumor core (NRC/NET -label1). Similar to BraTS2019, the task of BraTS2021 is to segment three overlap regions: enhancing tumor (ET, label 4), tumor core (TC, the composition of label 1 and 4), and whole tumor (WT, the composition of label 1, 2, and 4).

pre-processing Similar to the pre-processing methods in [Peiris et al., 2021], we performed a min-max scaling operation followed by clipping intensity values to standardize all volumes and crop/padding the volumes to a fixed size of $128 \times 128 \times 128$ by removing the unnecessary background. No data augmentation technique was applied and no additional data was used in this study.

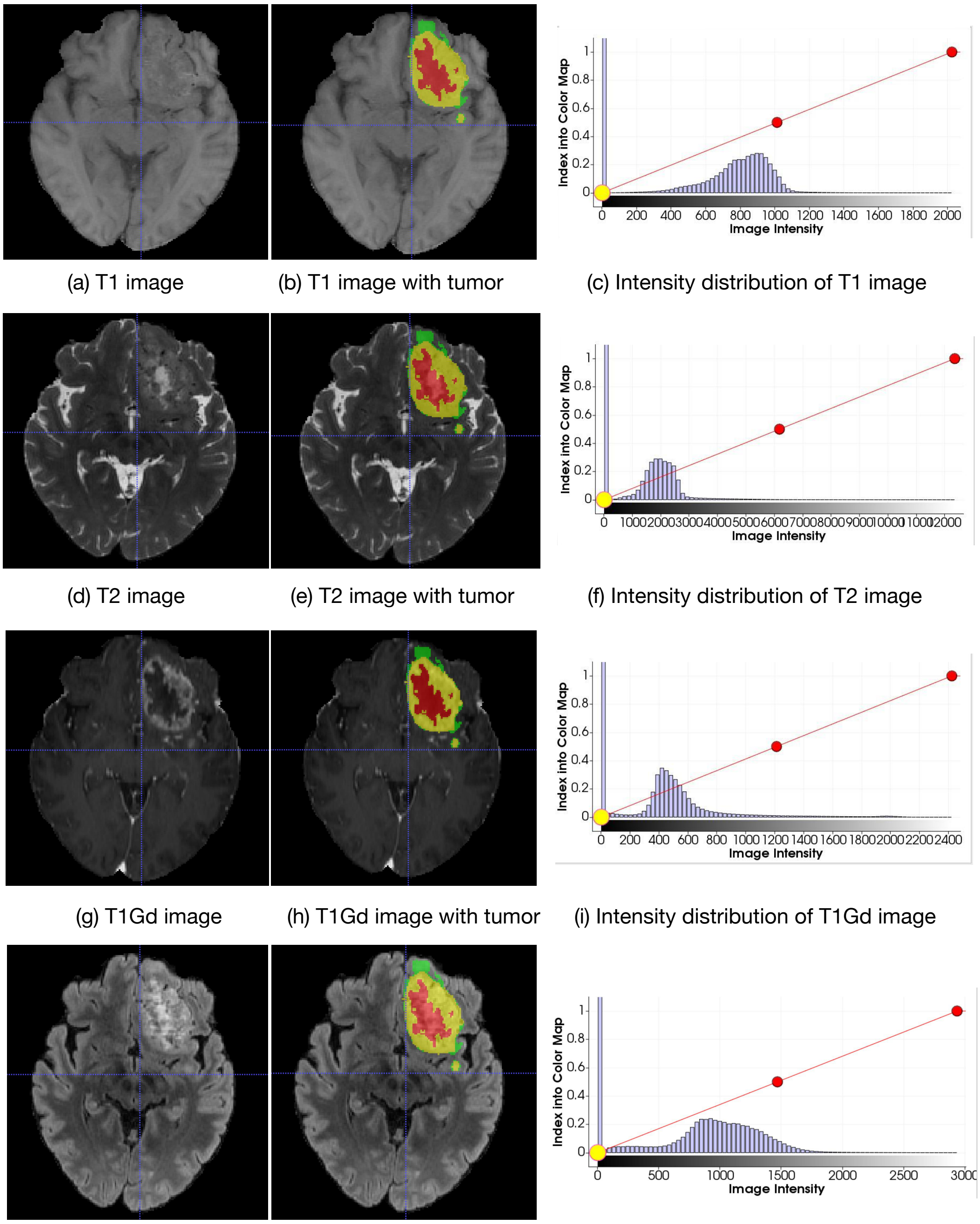
parameter initialization and learning The initial values of parameters (see (6.1b)) α_i and γ_i were set, respectively, to 0.5 and 0.01, and membership degrees u_{ic} were initialized randomly by drawing uniform random numbers and normalizing. The prototypes were initialized by the k-means clustering algorithm. For the multimodal evidence fusion module, the initial values of parameter β_c were set to 0.5. Each model was trained on the learning set with 100 epochs using the Adam optimization algorithm. The initial learning rate was set to 10^{-3}. The model with the best performance on the validation set was saved as the final model for testing[1].

The multimodal evidence fusion framework was implemented in Python with the PyTorch-based medical image framework MONAI, and was trained and tested on a desktop with a 2.20GHz Intel(R) Xeon(R) CPU E5-2698 v4 and a Tesla V100-SXM2 graphics card with 32 GB GPU memory.

6.3.2 *Segmentation accuracy*

We used the Dice Score (see (1.1)) and the Hausdorff Distance (HD) (see (1.4)) as our evaluation metrics to compare the segmentation accuracy with the baseline methods and state-of-the-art methods. For each patient, we separately computed these two indices for the three classes and then averaged indices over the patients, following a similar evaluation strategy as in [Peiris et al., 2021]. We use Residual-UNet (see Section 1.3.1) and nnUNet (see Section 1.3.3) as feature extraction modules to construct two models, named MMEF-UNet and MMEF-nnUNet, respectively. We compared our models with two recent transformer-based models (nnFormer

[1] The code is available at https://github.com/iWeisskohl.



(a) T1 image (b) T1 image with tumor (c) Intensity distribution of T1 image

(d) T2 image (e) T2 image with tumor (f) Intensity distribution of T2 image

(g) T1Gd image (h) T1Gd image with tumor (i) Intensity distribution of T1Gd image

(j) Flair image (k) Flair image with tumor (o) Intensity distribution of Flair image

Figure 6.2: Examples of a patient with the brain tumor in four modalities: T1, post-contrast T1-weighted (T1Gd), T2, FLAIR. The three tumor classes: enhancing core (yellow), necrotic/cystic core (red), and edema/invasion (green). The first, second and third columns show, respectively, the original image, the image with the tumor mask in three classes, and the intensity histogram of the image.

Table 6.1: Segmentation results on the BraTS 2021 test set. The best result is shown in bold, and the second best is underlined.

Methods	Dice score				HD			
	ET	TC	WT	Mean	ET	TC	WT	Mean
UNet [Çiçek et al., 2016]	83.39	86.28	89.59	86.42	11.49	6.18	6.15	7.94
VNet [Milletari et al., 2016]	81.04	84.71	90.32	85.36	17.20	7.48	7.53	10.73
nnFormer [Zhou et al., 2021]	82.83	86.48	90.37	86.56	11.66	7.89	8.00	9.18
VT-UNet [Peiris et al., 2021]	85.59	87.41	91.20	88.07	**10.03**	6.29	6.23	7.52
Residual-UNet [Kerfoot et al., 2018]	85.07	87.61	89.78	87.48	11.76	6.14	6.31	8.07
nnUNet [Isensee et al., 2018]	87.12	**90.31**	91.47	89.68	12.46	11.04	5.97	9.82
MMEF-UNet (Ours)	86.96	87.46	90.68	88.36	10.20	**6.07**	5.29	**7.18**
MMEF-nnUNet (Ours)	**87.26**	90.05	**92.83**	**90.05**	10.09	9.68	**5.10**	8.29

[Zhou et al., 2021], VT-UNet [Peiris et al., 2021]), two classical CNN-based methods (UNet [Çiçek et al., 2016], V-Net [Milletari et al., 2016]), and the two baseline models. The quantitative results are reported in Table 6.1. Our models outperform the two classical CNN-based models and two recent transformer-based methods in terms of the Dice score; the best result is obtained by MMEF-nnUNet according to this criterion. In contrast, MMEF-UNet achieves the lowest HD.

visualized segmentation results Figure 6.3 shows the segmentation results of Residual-UNet with the inputs of four concatenated modalities and MMEF-UNet with the inputs of four separate modalities. Our model locates and segments brain tumor subregions precisely, especially at the tumor boundary where uncertainty is high.

6.3.3 *Segmentation reliability*

A very important question related to the design of a multimodal medical image segmentation model is the reliability of its fusion results. Although all the recent publications report improved segmentation performance as the result of merging multimodal medical images into deep neural networks, the reliability of the results obtained with such fusion processes has seldom been investigated. There are two approaches to measuring the segmentation reliability. One is to test the reliability of the model by calculating the expected calibration error (ECE, see (5.10)). The other is to calculate the reliability of source information [Kobayashi et al., 1999]. In this chapter, the reliability of source information is represented by the reliability coefficients β we mentioned in (6.2).

model's reliability To test the reliability of our models, we computed the ECE for the two baseline models and ours. We obtain ECE values of 2.35 % and 2.04%, respectively, for Residual-UNet and MMEF-UNet, against ECE values of 4.46% and 4.05%, respectively, for nnUNet

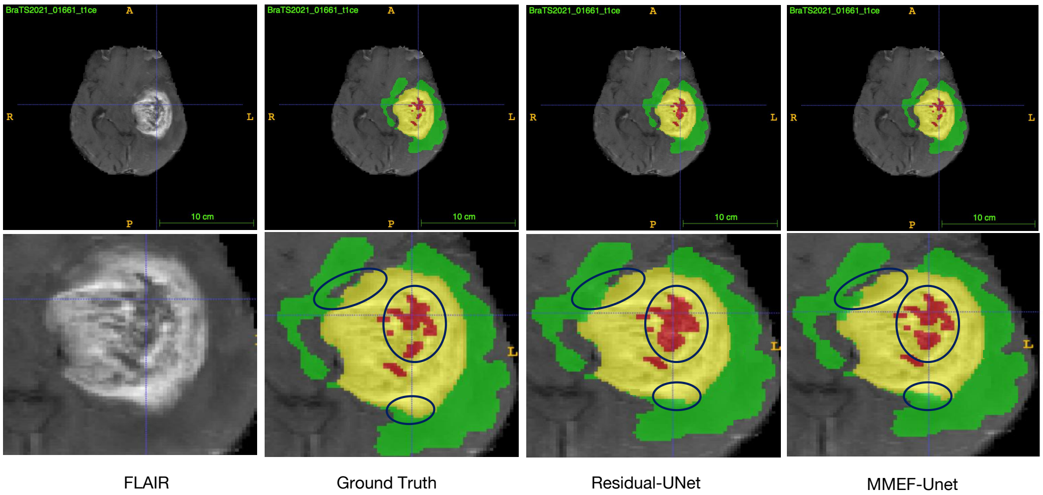


Figure 6.3: Visualization of the results on BraTs2021 data. The first and the second row are, respectively, the whole brain region and the tumor subregions (the main differences are marked in blue circles). The three columns correspond, from left to right, to the FLAIR image, the ground truth, and the segmentation results obtained by Residual-UNet and MMEF-UNet. The green, yellow and red represent the ET, ED, and NRC/NET, respectively.

and MMEF-nnUNet. Since the probabilities computed by our models are better calibrated, the segmentation results of our models are more reliable.

Table 6.2: Dice score and ECE on the BraTs2021 test set.

Methods	Dice score	ECE (%)
Residual-UNet	86.42	2.35
nnUNet	89.68	4.46
MMEF-UNet	88.36	2.04
MMEF-nnUNet	90.05	4.05

INTERPRETATION OF RELIABILITY COEFFICIENTS Table 6.3 shows the learned reliability coefficients β (see (6.2)) for the four modalities with three different classes. The evidence from the T1Gd modality is reliable for ET, ED, and NRC classes with the highest reliability value (close to 1). In contrast, the evidence from the FLAIR modality is only reliable for the ED class, with a reliability value coefficient to 0.86. The reliability coefficient of the T2 modality is only around 0.4 for the three classes, which means the T2 modality has a limited contribution to the final decision. The evidence from the T1 modality is less reliable for the three classes compared to the evidence of the other three modalities. These reliability

Table 6.3: Estimated reliability coefficient β (after training) for classes ET, ED and NRC/NET and the four modalities. Higher values correspond to greater contribution to the segmentation.

β	ET	ED	NRC/NET
T1Gd	0.9996	0.9726	0.9998
T1	0.4900	0.0401	0.2655
T2	0.4814	0.3881	0.4563
FLAIR	0.0748	0.86207	0.07512

results are consistent with domain knowledge about these modalities reported in [Baid et al., 2021], i.e., ET is described by areas with both visually avid, as well as faint, enhancement on T1Gd, the appearance of NRC/NET is hypointense on T1Gd. That is why its reliability coefficient is high for these areas. ED is defined by the abnormal hyperintense signal envelope on the FLAIR volumes. This interpretation of the reliability coefficients offers an effective way to explain the segmentation results to physicians and patients.

6.3.4 *Ablation analysis*

We also investigated the contribution of each module component to the performance of the framework. Table 6.4 highlights the importance of introducing the evidential segmentation and multimodal evidence fusion modules. Residual-UNet is the baseline model that uses the softmax transformation to map feature vectors into probabilities. Compared to Residual-UNet, Residual-UNet-ES uses the ES module instead of softmax that map feature vectors into mass functions. MMEF-UNet, our final proposal, fuses the four single-modality outputs from Residual-UNet-ES with the MMEF module.

illustrative results with single-modality inputs (residual-unet vs. residual-unet-es) Table 6.4 shows the performance of segmentation results with and without the evidential segmentation module when only single-modality MR images are used. Compared to the baseline method Residual-UNet, our proposal, which plugs the evidential segmentation module after Residual-UNet, improves the segmentation performance based on single inputs. For example, compared with Residual-UNet, Residual-UNet-ES has an increase in 1.38% and 0.62% of Dice score on T1Gd and FLAIR modality, respectively. Furthermore, the introduction of evidential segmentation decreases the ECE on four modalities, which means the segmentation results with the evidential segmentation module are more reliable.

Table 6.4: Segmentation results on BraTS 2021 validation set(↑ means higher is better).

Methods	Input	Dice score in single class			Mean Dice	ECE(%)
	Modality	ET	ED	NRC/NET		
Residual-UNet	T1Gd	79.55	61.16	71.26	71.30	3.67
Residual-UNet-ES	T1Gd	80.50 ↑	62.95↑	73.39↑	72.68↑	3.24↑
Residual-UNet	T1	44.32↑	50.58↑	44.85↑	48.97↑	4.27
Residual-UNet-ES	T1	44.41	56.08	44.06	48.55	3.33↑
Residual-UNet	T2	45.56	65.38	46.06	53.29	3.69
Residual-UNet-ES	T2	46.37↑	66.36↑	46.98↑	54.06↑	3.18 ↑
Residual-UNet	FLAIR	43.83	70.85	39.99	52.54	3.45
Residual-UNet-ES	FLAIR	45.45↑	71.58↑	40.12↑	53.16 ↑	3.43↑
Residual-UNet	T1Gd,T1 T2, FLAIR	86.06	81.68	77.07	81.60	2.35
MMEF-UNet	T1Gd,T1 T2, FLAIR	86.46↑	83.79↑	77.50↑	82.57↑	2.04 ↑

ILLUSTRATIVE RESULTS WITH DIFFERENT FUSION METHODS (CONCATENATING FUSION VS. MULTIMODAL EVIDENCE FUSION) We also compared the performance of fusion methods with image concatenation and multimodal evidence fusion. As we can see from Table 6.4, compared to any single-modality input, the fusion of multimodal medical images increases the Dice score by a significant amount. Furthermore, the use of the multimodal evidence fusion module improves the performance in terms of segmentation accuracy (Dice score) and reliability (ECE) in the subregions of ET, ED and NRC/NET compared to the input-level fusion that concatenates T1, T1Gd, T2, and FLAIR images as one single input. The above results demonstrate the effectiveness of our evidence fusion models.

6.4 CONCLUSION

Based on BFT, a multimodal evidence fusion framework considering segmentation uncertainty and source reliability has been proposed for multi-MRI brain tumor segmentation. The originality of our method is that the evidential segmentation module performs both tumor segmentation and uncertainty quantification, and the multimodal evidence fusion module carries out multimodal evidence fusion with contextual discounting and Dempster's rule.

This work is the first to implement contextual discounting for multimodal information fusion with BFT and deep neural networks. The contextual discounting operation allows us to take into account the uncertainty of the different sources of information directly, and it reveals the reliability of different modalities in different contexts as well. Our method can be considered as decision-level fusion and can be used together with any state-of-the-art feature extraction module to achieve better performance.

Some limitations of this work remain in the computation cost. We treat single modality images as independent inputs using independent feature extraction and evidential segmentation modules, which introduces additional computation costs compared to image concatenation methods (e.g., the FLOPs and parameter numbers are equal to 280.07G and 76.85M for UNet-MMEF, against 73.32G, and 19.21M for Residual-UNet). In future research, the refinement of the framework to improve segmentation performance and reduce its complexity will be considered.

Chapter 7

Conclusion and perspective

7.1 conclusion

The issue of uncertainty and reliability of machine learning methods has recently come to the forefront. It has incited researchers in the medical image segmentation domain to study both accurate and reliable segmentation methods [Kwon et al., 2020; Mehrtash et al., 2020; Ghoshal et al., 2021; Zou et al., 2022]. In this thesis, we studied on medical image segmentation with a specific focus on segmentation uncertainty and reliability under the framework of BFT and deep neural networks. In particular, we first briefly introduced medical image segmentation in Chapter 1 by describing the context of medical image segmentation, the deep learning medical segmentation methods and three baseline deep segmentation models. In Chapter 2, we then summarized the main concepts of BFT, i.e., representation of evidence, Dempster's combination rule, discounting, decision-making and the methods to generate mass functions. In Chapter 3, we introduced the BFT-based medical image segmentation methods in detail by grouping them according to the number of input modalities and classifiers used to generate mass functions, and we showed how unreliable or conflicting sources of information can be combined to reach reliable fusion results.

Our three main technique contributions have been presented in Chapters 4, 5 and 6. In Chapter 4, we have described a semi-supervised medical image segmentation model. The main idea is to introduce constraints for unlabelled data by generating pseudo labels and decrease the uncertainty caused by the lack of labels by fusing evidence from the probability distribution and evidence distribution. In Chapter 5, we introduced two evidential classifiers for uncertainty quantification and showed how they can be used with deep segmentation models to improve segmentation accuracy and reliability. In Chapter 6, we presented a multimodal medical image segmentation model that learns the reliability of each modality image when segmenting different types of tumors. We showed that using contextual discounting when fusing evidence from different modalities allows us to reach more reliable and explainable results. Our experimental results show that in many applications, BFT-based methods have the potential to outperform probability-based methods by modeling information more effectively and combining multiple piece of evidence at different stages.

7.2 FUTURE WORK

Despite the advantages of BFT for medical image segmentation, existing methods still have limitations that need to be addressed. Some directions for further research are discussed below.

First, acquiring large amounts of labeled training data is particularly challenging for medical image segmentation tasks and has become the bottleneck of learning-based segmentation performance. The successful application of unsupervised basic belief assignment methods to medical image segmentation points to a new direction to address the problem of lack of annotated training data. In this thesis, we have applied semi-supervised learning with BFT to improve the segmentation accuracy when only partial training data are labeled (see Chapter 4). To our best knowledge, there is no published paper dealing with the combination of unsupervised basic belief assignment methods with deep learning. The neural-network-based evidential clustering method described in [Denœux, 2021] and the EKNN rule with partially supervised learning introduced in [Denœux et al., 2019] are two steps in these directions. These works provide insights into using unsupervised or semi-supervised learning to quantify segmentation uncertainty with unannotated or partially annotated data sets. Future research could study the combination of unsupervised basic belief assignment methods with deep learning to overcome the annotation limitation.

Second, most of the existing BFT-based medical image segmentation methods still use low-level features and do not fully exploit the advantages of deep learning. Combining BFT with deep segmentation networks should allow us to develop accurate and reliable segmentation models, particularly in medical image segmentation tasks for which medical knowledge is available and can be modeled by mass functions. In this thesis, we have revealed the similarity of ENN and RBF network when acting as an evidential classifier and integrated both classifiers within a deep segmentation network to improve the segmentation accuracy and reliability (see Chapter 5). We believe that more promising results will be obtained by blending BFT with the existing powerful deep medical image segmentation models, e.g., transformer [Shamshad et al., 2022] and diffusion models [Kazerouni et al., 2022].

Third, even though the BFT-based fusion methods have considered the conflict of multiple sources of evidence, a problem remains that multimodal medical images may have different degrees of reliability when segmenting different type of tumors or organs. Multimodal medical image fusion may fail if some sources of information are unreliable or partially reliable for performing various segmentation tasks. In this thesis, we have proposed a method to learn reliability coefficients conditionally on different segmentation tasks in a deep neural network (see Chapter 6). Another interesting research direction, accordingly, is to estimate task-specific reliability and enhance the explainability of deep evidential neural networks using contextual discounting for multimodal or cross-modal medical image segmentation tasks.

We hope our work will increase awareness of the challenges of existing BFT-based medical image segmentation methods, call for future contributions to bridge the gap between experimental performance and clinical application, and develop accurate, reliable, and explainable deep segmentation models.

Bibliography

Agravat, Rupal R and Mehul S Raval (2019). "Brain tumor segmentation and survival prediction". In: *International MICCAI Brainlesion Workshop*. Springer. Shenzhen, China, pp. 338–348 (cited on p. 73).

Appriou, A. (1999). "Multisensor signal processing in the framework of the theory of evidence". In: *Application of mathematical signal processing techniques to mission systems* 32 (cited on pp. 28, 29).

— (2005). "01-Approche générique de la gestion de l'incertain dans les processus de fusion multisenseur". In: *Traitement du signal* (cited on pp. 28, 29).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (cited on pp. 1, 15).

Bai, C., L. Huang, X. Pan, J. Zheng, and S. Chen (2018). "Optimization of deep convolutional neural network for large scale image retrieval". In: *Neurocomputing* 303, pp. 60–67 (cited on p. 13).

Baid, Ujjwal, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. (2021). "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification". In: *arXiv preprint arXiv:2107.02314* (cited on pp. 109, 114).

Bakas, Spyridon, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos (2017). "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1, pp. 1–13 (cited on p. 65).

Bakas, Spyridon, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. (2018). "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (cited on p. 65).

Barhoumi, W., S. Dhahbi, and E. Zagrouba (2007). "A collaborative system for pigmented skin lesions malignancy tracking". In: *2007 IEEE International Workshop on Imaging Systems and Techniques*. IEEE. Cracovia, Poland, pp. 1–6 (cited on pp. 50, 52).

Batenburg K., J. and J. Sijbers (2009). "Adaptive thresholding of tomograms by projection distance minimization". In: *Pattern Recognition* 42.10, pp. 2297–2305 (cited on pp. 1, 13).

Benjamini, Yoav and Daniel Yekutieli (2001). “The control of the false discovery rate in multiple testing under dependency”. In: *Annals of statistics*, pp. 1165–1188 (cited on p. 97).

Bezdek, J.C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media (cited on p. 56).

Bloch I.and Maître, H. (1995). “Fuzzy mathematical morphologies: a comparative study”. In: *Pattern recognition* 28.9, pp. 1341–1387 (cited on p. 44).

Bloch, I. (1996). “Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account”. In: *Pattern Recognition Letters* 17.8, pp. 905–919 (cited on pp. 44, 45).

— (2008). “Defining belief functions using mathematical morphology–application to image fusion under imprecision”. In: *International journal of approximate reasoning* 48.2, pp. 437–465 (cited on pp. 44, 45).

Bowyer, K., D. Kopans, W.P. Kegelmeyer, R. Moore, M. Sallam, K. Chang, and K. Woods (1996). “The digital database for screening mammography”. In: *Third international workshop on digital mammography*. Vol. 58. Mammography, Chicago, p. 27 (cited on p. 56).

Broadhouse, Kathryn (2020). “The physics of MRI and how we use it to reveal the mysteries of the mind”. In: *Frontiers for Young Minds* 7 (cited on p. 11).

Caliva, Francesco, Claudia Iriondo, Alejandro Morales Martinez, Sharmila Majumdar, and Valentina Pedoia (2019). “Distance map loss penalty term for semantic segmentation”. In: *arXiv preprint arXiv:1908.03679* (cited on p. 15).

Cao, Hu, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang (2021). “Swin-unet: Unet-like pure transformer for medical image segmentation”. In: *arXiv preprint arXiv:2105.05537* (cited on p. 15).

Capelle, A.-S., O. Colot, and C. Fernandez-Maloigne (2002). “Segmentation of multi-modality MR images by means of Evidence Theory for 3D reconstruction of brain tumors”. In: *Proceedings. International Conference on Image Processing*. Vol. 2. IEEE. New York, USA, pp. II–II (cited on pp. 50, 51).

Capelle, A-S., O. Colot, and C. Fernandez-Maloigne (2004). “Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information”. In: *Information Fusion* 5.3, pp. 203–216 (cited on pp. 50, 51).

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). “End-to-end object detection with transformers”. In: *European conference on computer vision*. Springer, pp. 213–229 (cited on p. 15).

Chaabane, S.-Ben, F. Sayadi M.and Fnaiech, E. Brassart, and F. Betin (2011). "A new method for the estimation of mass functions in the Dempster-Shafer's evidence theory: application to colour image segmentation". In: *Circuits, Systems, and Signal Processing* 30.1, pp. 55–71 (cited on pp. 45, 56).

Chaabane, S.B., F. Fnaiech, M. Sayadi, and E. Brassart (2009). "Relevance of the Dempster-Shafer evidence theory for image segmentation". In: *3rd International Conference on Signals, Circuits and Systems (SCS)*. IEEE. Medenine, Tunisia, pp. 1–4 (cited on pp. 45, 47, 56).

Chan, T.-F and L.-A Vese (2001). "Active contours without edges". In: *IEEE Transactions on image processing* 10.2, pp. 266–277 (cited on p. 52).

Chan, T.F., B.Y. Sandberg, and L.A. Vese (2000). "Active contours without edges for vector-valued images". In: *Journal of Visual Communication and Image Representation* 11.2, pp. 130–141 (cited on p. 42).

Chen, Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li (2019a). "3D dilated multi-fiber network for real-time brain tumor segmentation in MRI". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. Shenzhen, China, pp. 184–192 (cited on p. 73).

Chen, Shuai, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne (2019b). "Multi-task attention-based semi-supervised learning for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. Shenzhen, China, pp. 457–465 (cited on p. 75).

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR, pp. 1597–1607 (cited on p. 16).

Chen, X., D. Zou, Q. Zhao, and P. Tan (2012). "Manifold preserving edit propagation". In: *ACM Transactions on Graphics (TOG)* 31.6, pp. 1–7 (cited on pp. 32, 35).

Chen, Yunpeng, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng (2018). "Multi-fiber networks for video recognition". In: *Proceedings of the european conference on computer vision (ECCV)*. Munich, Germany, pp. 352–367 (cited on pp. 18, 19).

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. Athens, Greece, pp. 424–432 (cited on pp. 15, 112).

Ciresan, D., A. Giusti, L. M Gambardella, and J. Schmidhuber (2012). "Deep neural networks segment neuronal membranes in electron microscopy images". In: *Advances in Neural Information Processing Systems*. Lake Tahoe, USA, pp. 2843–2851 (cited on p. 13).

Conover, W J and R L Iman (Feb. 1979). "Multiple-comparisons procedures. Informal report". In: DOI: 10.2172/6057803. URL: https://www.osti.gov/biblio/6057803 (cited on p. 97).

Dempster, A.-P (1967). "Upper and lower probability inferences based on a sample from a finite univariate population". In: *Biometrika* 54.3-4, pp. 515–528 (cited on pp. 2, 23).

Denœux, T. (1995). "A k-nearest neighbor classification rule based on Dempster-Shafer theory". In: *IEEE Transactions on Systems, Man, and Cybernetics* 25.5, pp. 804–813. DOI: 10.1109/21.376493 (cited on pp. 23, 29).

— (2000). "A neural network classifier based on Dempster-Shafer theory". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30.2, pp. 131–150 (cited on pp. 23, 29–31, 79, 82, 102).

— (2008). "Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence". In: *Artificial Intelligence* 172.2-3, pp. 234–264 (cited on p. 23).

— (2016). "40 years of Dempster-Shafer theory". In: *International Journal of Approximate Reasoning* 79.C, pp. 1–6 (cited on p. 24).

— (2019a). "Decision-making with belief functions: A review". In: *International Journal of Approximate Reasoning* 109, pp. 87–110 (cited on pp. 28, 29, 39).

— (2019b). "Logistic regression, neural networks and Dempster-Shafer theory: A new perspective". In: *Knowledge-Based Systems* 176, pp. 54–67 (cited on pp. 79–81).

— (2021). "NN-EVCLUS: Neural Network-based Evidential Clustering". In: *Information Sciences* 572, pp. 297–330 (cited on p. 118).

Denœux, T. and M.-H. Masson (2004). "EVCLUS: evidential clustering of proximity data". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34.1, pp. 95–109 (cited on pp. 23, 31, 34).

Denœux, T., O. Kanjanatarakul, and S. Sriboonchitta (2019). "A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning". In: *International Journal of Approximate Reasoning* 113, pp. 287–302 (cited on pp. 27, 30, 118).

Denœux, Thierry, Didier Dubois, and Henri Prade (2020). "Representations of uncertainty in artificial intelligence: Probability and possibility". In: *A Guided Tour of Artificial Intelligence Research*. Springer, pp. 69–117 (cited on p. 2).

Der Kiureghian, Armen and Ove Ditlevsen (2009). "Aleatory or epistemic? Does it matter?" In: *Structural safety* 31.2, pp. 105–112 (cited on p. 1).

Derraz F.and Pinti, A., M. Boussahla, L. Peyrodie, and H. Toumi (2013a). "Image Segmentation Using Active Contours and Evidential Distance". In: *Iberoamerican Congress on Pattern Recognition*. Springer. Havana, Cuba, pp. 472–479 (cited on pp. 41, 42, 48).

Derraz, F., M. Boussahla, and L. Peyrodie (2013b). “Globally Segmentation Using Active Contours and Belief Function”. In: *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer. Poznań, Poland, pp. 546–554 (cited on pp. 41, 42, 48).

Derraz, F., A. Pinti, L. Peyrodie, M. Bousahla, and H. Toumi (2015). “Joint variational segmentation of CT/PET data using non-local active contours and belief functions”. In: *Pattern Recognition and Image Analysis* 25.3, pp. 407–412 (cited on pp. 45, 48).

Dinno, Alexis and Maintainer Alexis Dinno (2017). “Package conover test”. In: *Conover-Iman test of multiple comparisons using rank* (cited on p. 97).

Duan, R. and Y.-H. Guan (2008). “Multi-threshold value segmentation approach for medical images [J]”. In: *Journal of Computer Applications* 28.S2, pp. 196–197 (cited on p. 52).

Dubois, D. and H. Prade (1988). “Representation and combination of uncertainty with belief functions and possibility measures”. In: *Computational intelligence* 4.3, pp. 244–264 (cited on p. 23).

— (2012). *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science & Business Media (cited on p. 24).

Dunn, J.C. (1973). “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”. In: *Journal of Cybernetics* 3.3, pp. 32–57. DOI: 10.1080/01969727308546046 (cited on p. 31).

Gal, Yarin and Zoubin Ghahramani (2016). “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR, pp. 1050–1059 (cited on pp. 2, 99).

Gautier, L., A. Taleb-Ahmed, M. Rombaut, J.-G. Postaire, and H. Leclet (2000). “Belief function in low level data fusion: application in MRI images of vertebra”. In: *Proceedings of the Third International Conference on Information Fusion*. Vol. 1. IEEE. Paris, France, TUC5–3 (cited on pp. 54, 56).

George, K., S. Faziludeen, and P. Sankaran (2020). “Breast cancer detection from biopsy images using nucleus guided transfer learning and belief based fusion”. In: *Computers in Biology and Medicine* 124, p. 103954 (cited on pp. 50, 53, 54).

Gerig, G., D. Welti, RG. Guttmann C., CF. Colchester A., and G. Székely (2000). “Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data”. In: *Medical Image Analysis* 4.1, pp. 31–42 (cited on p. 41).

Ghasemi, J., M. R. K. Mollaei, R. Ghaderi, and A. Hojjatoleslami (2012). “Brain tissue segmentation based on spatial information fusion by Dempster-Shafer theory”. In: *Journal of Zhejiang University SCIENCE C* 13.7, pp. 520–533 (cited on pp. 31, 33, 45).

Ghasemi, J., R. Ghaderi, MR-K. Mollaei, and SA. Hojjatoleslami (2013). "A novel fuzzy Dempster-Shafer inference system for brain MRI segmentation". In: *Information Sciences* 223, pp. 205–220 (cited on p. 45).

Ghesu, F.C., B. Georgescu, A. Mansoor, Y. Yoo, E. Gibson, R.S. Vishwanath, A. Balachandran, J.M. Balter, Y. Cao, and R. Singh (2021). "Quantifying and leveraging predictive uncertainty for medical image assessment". In: *Medical Image Analysis* 68, p. 101855 (cited on p. 3).

Ghoshal, Biraja, Allan Tucker, Bal Sanghera, and Wai Lup Wong (2021). "Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection". In: *Computational Intelligence* 37.2, pp. 701–734 (cited on p. 117).

Goldstein, T. and S. Osher (2009). "The split Bregman method for L1-regularized problems". In: *SIAM journal on imaging sciences* 2.2, pp. 323–343 (cited on p. 42).

Guan, Y.-H., Y. Luo, T. Yang, L. Qiu, and J. Li (2011). "Study on the Application of MRF and Fuzzy Clustering as well as the DS Theory to Image Fusion Segmentation of the Human Brain". In: *4th International Conference on Intelligent Networks and Intelligent Systems*. IEEE. Washington, USA, pp. 57–60 (cited on pp. 50, 52).

Guo, C., G. Pleiss, Y. Sun, and K.Q. Weinberger (2017). "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1321–1330 (cited on p. 89).

Han, Kai, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang (2021). "Transformer in transformer". In: *Advances in Neural Information Processing Systems* 34, pp. 15908–15919 (cited on p. 15).

Hanin, Boris and David Rolnick (2018). "How to start training: The effect of initialization and architecture". In: *arXiv preprint arXiv:1803.01719* (cited on p. 68).

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik (2015). "Hypercolumns for object segmentation and fine-grained localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, USA, pp. 447–456 (cited on p. 13).

Harrabi, R. and E. B. Braiek (2012). "Color image segmentation using multi-level thresholding approach and data fusion techniques: application in the breast cancer cells images". In: *EURASIP Journal on Image and Video Processing* 2012.1, p. 11 (cited on pp. 45, 47).

Hatamizadeh, Ali, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu (2022a). "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images". In: *arXiv preprint arXiv:2201.01266* (cited on p. 15).

Hatamizadeh, Ali, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu (2022b).

"Unetr: Transformers for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584 (cited on p. 15).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034 (cited on p. 86).

— (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV, USA, pp. 770–778 (cited on p. 19).

Hinton, Geoffrey E and Drew Van Camp (1993). "Keeping the neural networks simple by minimizing the description length of the weights". In: *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13 (cited on p. 2).

Hora, Stephen C (1996). "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management". In: *Reliability Engineering & System Safety* 54.2-3, pp. 217–223 (cited on p. 1).

Hryniowski, Andrew and Alexander Wong (2019). "DeepLABNet: End-to-end learning of deep radial basis networks". In: *Journal of Computational Vision and Imaging Systems* 5.1, pp. 1–1 (cited on p. 102).

Huang, L., S. Ruan, and T. Denœux (2021a). "Belief function-based semi-supervised learning for brain tumor segmentation". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 160–164 (cited on pp. 29, 50, 53, 54).

Huang, L., T. Denœux, D. Tonnelet, P. Decazes, and S. Ruan (2021b). "Deep PET/CT fusion with Dempster-Shafer theory for lymphoma segmentation". In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 30–39 (cited on pp. 54, 56).

Huang, L., S. Ruan, P. Decazes, and T. Denœux (2021c). "Evidential segmentation of 3D PET/CT images". In: *International Conference on Belief Functions*. Springer, pp. 159–167 (cited on pp. 41–43).

Huang, L., T. Denœux, P. Vera, and S. Ruan (2022a). "Evidence Fusion with Contextual Discounting for Multi-modality Medical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, pp. 401–411. ISBN: 978-3-031-16443-9 (cited on pp. 54, 57).

Huang, L., S. Ruan, P. Decazes, and T. Denœux (2022b). "Lymphoma segmentation from 3D PET-CT images using a deep evidential network". In: *International Journal of Approximate Reasoning* 149, pp. 39–60. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2022.06.007 (cited on pp. 29, 41–43).

Hüllermeier, Eyke and Willem Waegeman (2021). “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine Learning* 110.3, pp. 457–506 (cited on pp. 1, 77).

Isensee, F., J. Petersen, . Klein, and D. Zimmerer (2018). “nnU-Net: Self-adapting framework for u-net-based medical image segmentation”. In: *arXiv preprint arXiv:1809.10486* (cited on pp. 1, 97, 98, 112).

Isensee, Fabian, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein (2021). “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2, pp. 203–211 (cited on pp. 15, 18, 20, 21).

Islam, Mobarakol, VS Vibashan, V Jeya Maria Jose, Navodini Wijethilake, Uppal Utkarsh, and Hongliang Ren (2019). “Brain tumor segmentation and survival prediction using 3D attention UNet”. In: *International MICCAI Brainlesion Workshop*. Springer. Shenzhen, China, pp. 262–272 (cited on p. 73).

Jhanwar, Yuliya S and David J Straus (2006). “The role of PET in lymphoma”. In: *Journal of Nuclear Medicine* 47.8, pp. 1326–1334 (cited on p. 78).

Jiang, Zeyu, Changxing Ding, Minfeng Liu, and Dacheng Tao (2019). “Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task”. In: *International MICCAI Brainlesion Workshop*. Springer. Shenzhen, China, pp. 231–241 (cited on pp. 71, 73).

Jungo, Alain, Fabian Balsiger, and Mauricio Reyes (2020). “Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation”. In: *Frontiers in neuroscience*, p. 282 (cited on p. 90).

Kamnitsas, K., W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, and D. Rueckert (2017). “Ensembles of multiple models and architectures for robust brain tumour segmentation”. In: *International MICCAI Brainlesion Workshop*. Springer. Cham, pp. 450–462 (cited on pp. 43, 106).

Karimi, Davood and Septimiu E Salcudean (2019). “Reducing the hausdorff distance in medical image segmentation with convolutional neural networks”. In: *IEEE Transactions on medical imaging* 39.2, pp. 499–513 (cited on p. 16).

Kass, M. and D. Witkin A.and Terzopoulos (1988). “Snakes: Active contour models”. In: *International journal of computer vision* 1.4, pp. 321–331 (cited on p. 52).

Kazerouni, Amirhossein, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof (2022). “Diffusion Models for Medical Image Analysis: A Comprehensive Survey”. In: *arXiv preprint arXiv:2211.07804* (cited on p. 118).

Kerfoot, Eric, James Clough, Ilkay Oksuz, Jack Lee, Andrew P King, and Julia A Schnabel (2018). "Left-ventricle quantification using residual U-Net". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, pp. 371–380 (cited on pp. 18, 98, 112).

Ketout, H., J. Gu, and G. Horne (2012). "Improved Dempster and Shafer theory to fuse region and edge based level set for endocardial contour detection". In: *Proceedings of the 10th World Congress on Intelligent Control and Automation*. IEEE. Beijing, China, pp. 5013–5018 (cited on pp. 50, 52).

Kimmel, R. (2003). "Fast edge integration". In: *Geometric Level Set Methods in Imaging, Vision, and Graphics*. New York, USA: Springer, pp. 59–77 (cited on pp. 1, 13).

Kleinbaum, D.G. and D. Hedeker (1996). "Logistic regression: a self-learning text". In: *Statistical Methods in Medical Research* 5.1, p. 103 (cited on pp. 1, 13).

Kobayashi, Futoshi, Fumihito Arai, and Toshio Fukuda (1999). "Sensor selection by reliability based on possibility measure". In: *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*. Vol. 4. IEEE, pp. 2614–2619 (cited on p. 112).

Krizhevsky, A., I. Sutskever, and G. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Lake Tahoe, USA, pp. 1097–1105 (cited on p. 13).

Kwon, Yongchan, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik (2020). "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation". In: *Computational Statistics & Data Analysis* 142, p. 106816 (cited on p. 117).

Lai, K.F. (1994). "Deformable contours: Modeling, extraction, detection and classification". PhD thesis. University of Wisconsin–Madison (cited on p. 56).

Lajili, R., K. Kalti, A. Touil, B. Solaiman, and N. E. B. Amara (2018). "Two-step evidential fusion approach for accurate breast region segmentation in mammograms". In: *IET Image Processing* 12.11, pp. 1972–1982 (cited on pp. 54, 56).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (cited on p. 2).

Lelandais, B., I. Gardin, L. Mouchard, P. Vera, and S. Ruan (2012). "Segmentation of biological target volumes on multi-tracer PET images based on information fusion for achieving dose painting in radiotherapy". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. Nice, France, pp. 545–552 (cited on pp. 45, 48).

Lelandais, B., S. Ruan, T. Denœux, P. Vera, and I. Gardin (2014). "Fusion of multi-tracer PET images for dose painting". In: *Medical image analysis* 18.7, pp. 1247–1259 (cited on pp. 42, 45, 48, 49).

Li, C., C. Xu, C. Gui, and M.-D Fox (2005). "Level set evolution without re-initialization: a new variational formulation". In: *IEEE computer society conference on computer vision and pattern recognition*. Vol. 1. IEEE. San Diego, California, pp. 430–436 (cited on p. 52).

Li, S.Z. (2009). *Markov random field modeling in image analysis*. Springer Science & Business Media (cited on p. 52).

Li, X., L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng (2019a). "Transformation Consistent Self-ensembling Model for Semi-supervised Medical Image Segmentation". In: *arXiv preprint arXiv:1903.00348* (cited on p. 61).

Li, Xiangyu, Gongning Luo, and Kuanquan Wang (2019b). "Multi-step cascaded networks for brain tumor segmentation". In: *International MICCAI Brainlesion Workshop*. Springer. Shenzhen, China, pp. 163–173 (cited on p. 73).

Lian, C., S. Ruan, T. Denœux, Y. Guo, and P. Vera (2017a). "Accurate tumor segmentation in FDG-PET images with guidance of complementary CT images". In: *International Conference on Image Processing (ICIP)*. IEEE. Beijing, China, pp. 4447–4451 (cited on pp. 45, 49).

Lian, C., S. Ruan, T. Denœux, H. Li, and P. Vera (2017b). "Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images". In: *IEEE Transactions on Biomedical Engineering* 65.1, pp. 21–30 (cited on pp. 41, 42, 49).

Lian, C., T. Ruan S.and Denœux, H. Li, and P. Vera (2017c). "Tumor delineation in FDG-PET images using a new evidential clustering algorithm with spatial regularization and adaptive distance metric". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. Melbourne, Australia, pp. 1177–1180 (cited on p. 41).

Lian, C., S. Ruan, T. Denœux, H. Li, and P. Vera (2018a). "Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions". In: *IEEE Transactions on Image Processing* 28.2, pp. 755–766 (cited on pp. 3, 45, 49).

Lian, Chunfeng, Hua Li, Pierre Vera, and Su Ruan (2018b). "Unsupervised co-segmentation of tumor in PET-CT images using belief functions based fusion". In: *2018 Ieee 15th International Symposium On Biomedical Imaging (Isbi 2018)*. IEEE, pp. 220–223 (cited on pp. 45, 49).

Lima, S.-A. and Md-R. Islam (2019). "A Modified Method for Brain MRI Segmentation Using Dempster-Shafer Theory". In: *22nd International Conference on Computer and Information Technology*. IEEE. Helsinki, Finland, pp. 1–6 (cited on pp. 45, 46).

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988 (cited on p. 15).

Long, J., E. Shelhamer, and T. Darrell (2015). “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, USA, pp. 3431–3440 (cited on p. 13).

Lowe, David G (1999). “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee, pp. 1150–1157 (cited on p. 13).

Lowekamp, Bradley C, David T Chen, Luis Ibáñez, and Daniel Blezek (2013). “The design of SimpleITK”. In: *Frontiers in neuroinformatics* 7, p. 45 (cited on p. 87).

Ma, Jun (2020). “Segmentation loss odyssey”. In: *arXiv preprint arXiv:2005.13449* (cited on p. 16).

MacKay, David JC (1992). “A practical Bayesian framework for backpropagation networks”. In: *Neural computation* 4.3, pp. 448–472 (cited on p. 2).

Maddox, Wesley J, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson (2019). “A simple baseline for bayesian uncertainty in deep learning”. In: *Advances in Neural Information Processing Systems* 32 (cited on p. 77).

Makni, N., N. Betrouni, and O. Colot (2014). “Introducing spatial neighbourhood in Evidential C-Means for segmentation of multi-source images: Application to prostate multi-parametric MRI”. In: *Information Fusion* 19, pp. 61–72 (cited on pp. 45, 46, 48).

Masson, M.-H. and T. Denœux (2008). “ECM: An evidential version of the fuzzy c-means algorithm”. In: *Pattern Recognition* 41.4, pp. 1384–1397 (cited on pp. 31, 34).

— (2009). “RECM: Relational evidential c-means algorithm”. In: *Pattern Recognition Letters* 30.11, pp. 1015–1026 (cited on p. 23).

Mehrtash, Alireza, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur (2020). “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation”. In: *IEEE transactions on medical imaging* 39.12, pp. 3868–3878 (cited on p. 117).

Mehta, Raghav, Thomas Christinck, Tanya Nair, Paul Lemaitre, Douglas Arnold, and Tal Arbel (2019). “Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*. Springer, pp. 23–32 (cited on p. 77).

Menze, Bjoern H, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. (2014). “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10, pp. 1993–2024 (cited on p. 65).

Mercier, D., B. Quost, and T. Denœux (2008). “Refined modeling of sensor reliability in the belief function framework using contextual discounting”. In: *Information fusion* 9.2, pp. 246–258 (cited on pp. 27, 108).

Mercier, David, Éric Lefevre, and François Delmotte (2012). "Belief functions contextual discounting and canonical decompositions". In: *International Journal of Approximate Reasoning* 53.2, pp. 146–158 (cited on p. 106).

Mian, Ajmal S, Mohammed Bennamoun, and Robyn Owens (2006). "Three-dimensional model-based object recognition and segmentation in cluttered scenes". In: *IEEE transactions on pattern analysis and machine intelligence* 28.10, pp. 1584–1601 (cited on p. 13).

Michailovich, O., Y. Rathi, and A. Tannenbaum (2007). "Image segmentation using active contours driven by the Bhattacharyya gradient flow". In: *IEEE Transactions on Image Processing* 16.11, pp. 2787–2801 (cited on p. 42).

Milletari, F., N. Navab, and S.-A. Ahmadi (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision*. IEEE, pp. 565–571 (cited on pp. 15, 97, 98, 112).

Min, Shaobo, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang (2019). "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 4578–4585 (cited on pp. 61, 75).

Mohamed, M. and B. Far (2012). "An enhanced threshold based technique for white blood cells nuclei automatic segmentation". In: *14th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE. Beijing, China, pp. 202–207 (cited on p. 42).

Mondal, A.K., J. Dolz, and C. Desrosiers (2018). "Few-shot 3D multi-modal medical image segmentation using generative adversarial learning". In: *arXiv preprint arXiv:1810.12241* (cited on p. 61).

Myronenko, Andriy (2018). "3D MRI brain tumor segmentation using autoencoder regularization". In: *International MICCAI Brainlesion Workshop*. Springer, pp. 311–320 (cited on pp. 1, 15, 97, 98).

Niemeijer, M., B. Van Ginneken, M.J. Cree, A. Mizutani, G. Quellec, C.I. Sánchez, B. Zhang, R. Hornero, M. Lamard, and C. Muramatsu (2009). "Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs". In: *IEEE transactions on medical imaging* 29.1, pp. 185–195 (cited on p. 42).

Oktay, O., J. Schlemper, L.L. Folgoc, M. Lee M.and Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y Hammerla, and B. Kainz (2018). "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (cited on p. 15).

Onoma, D.P., S. Ruan, S. Thureau, L. Nkhali, R. Modzelewski, G.A. Monnehan, P. Vera, and I. Gardin (2014). "Segmentation of heterogeneous or small FDG PET positive tissue based on a 3D-locally adaptive random walk algorithm". In: *Computerized Medical Imaging and Graphics* 38.8, pp. 753–763 (cited on p. 13).

Peiris, H., M. Hayat, Z. Chen, G. Egan, and M. Harandi (2021). "A Volumetric Transformer for Accurate 3D Tumor Segmentation". In: *arXiv preprint arXiv:2111.13300* (cited on pp. 43, 105, 109, 110, 112).

Pelberg, Robert (2015). "Basic principles in computed tomography (ct)". In: *Cardiac CT angiography manual*. Springer, pp. 19–58 (cited on p. 9).

Pichon, Frédéric, David Mercier, Éric Lefevre, and François Delmotte (2016). "Proposition and learning of some belief function contextual correction mechanisms". In: *International Journal of Approximate Reasoning* 72, pp. 4–42 (cited on p. 106).

Quellec, G., M. Lamard, P.M. Josselin, G. Cazuguel, B. Cochener, and C. Roux (2008). "Optimal wavelet transform for the detection of microaneurysms in retina photographs". In: *IEEE transactions on medical imaging* 27.9, pp. 1230–1241 (cited on p. 42).

Quinonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence (2008). *Dataset shift in machine learning*. Mit Press (cited on p. 77).

Rahman, Md Atiqur and Yang Wang (2016). "Optimizing intersection-over-union in deep neural networks for image segmentation". In: *International symposium on visual computing*. Springer, pp. 234–244 (cited on p. 15).

Rogova, Galina L and Vincent Nimier (2004). "Reliability in information fusion: literature survey". In: *Proceedings of the seventh international conference on information fusion*. Vol. 2, pp. 1158–1165 (cited on p. 2).

Ronneberger, O., P. Fischer, and TU-net. Brox (2015). "Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich, Germany (cited on pp. 1, 13–15, 62, 97).

Rosten, Edward and Tom Drummond (2006). "Machine learning for high-speed corner detection". In: *European conference on computer vision*. Springer, pp. 430–443 (cited on p. 13).

Rousseau, Axel-Jan, Thijs Becker, Jeroen Bertels, Matthew B Blaschko, and Dirk Valkenborg (2021). "Post training uncertainty calibration of deep networks for medical image segmentation". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1052–1056 (cited on p. 90).

Rudroff, Thorsten, John H Kindred, and Kari K Kalliokoski (2015). "[18F]-FDG positron emission tomography—an established clinical tool opening a new window into exercise physiology". In: *Journal of Applied Physiology* 118.10, pp. 1181–1190 (cited on p. 10).

Rupprecht, Christian, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager (2017). "Learning in an uncertain world: Representing ambiguity through multiple hypotheses". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3591–3600 (cited on p. 2).

Safranek, R.J., S. Gottschlich, and A Kak (1990). "Evidence accumulation using binary frames of discernment for verification vision". In: *IEEE Transactions on Robotics and Automation* 6.4, pp. 405–417 (cited on pp. 31, 35).

Salehi, Seyed Sadegh Mohseni, Deniz Erdogmus, and Ali Gholipour (2017). "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International workshop on machine learning in medical imaging*. Springer, pp. 379–387 (cited on p. 15).

Salvador, S. and P. Chan (2004). "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms". In: *16th IEEE international conference on tools with artificial intelligence*. IEEE. Boca Raton, USA, pp. 576–584 (cited on pp. 1, 13).

Seyedhosseini, M., M. Sajjadi, and T. Tasdizen (2013). "Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. Sydney, Australia, pp. 2168–2175 (cited on p. 13).

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press (cited on pp. 2, 23, 25–29, 40, 106).

Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu (2022). "Transformers in medical imaging: A survey". In: *arXiv preprint arXiv:2201.09873* (cited on p. 118).

Smets, P. (1990). "The combination of evidence in the transferable belief model". In: *IEEE Transactions on pattern analysis and machine intelligence* 12.5, pp. 447–458 (cited on pp. 23, 24, 28).

Smets, Ph. and R. Kennes (1994). "The Transferable Belief Model". In: *Artificial Intelligence* 66, pp. 191–243 (cited on p. 27).

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, p. 25 (cited on pp. 1, 13).

Suckling, J.P. (1994). "The mammographic image analysis society digital mammogram database". In: *Digital Mammo*, pp. 375–386 (cited on p. 56).

Sudre, Carole H, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso (2017). "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*. Springer, pp. 240–248 (cited on pp. 15, 16).

Suh, D.Y., R.L. Eisner, R.M. Mersereau, and R.I. Pettigrew (1993). "Knowledge-based system for boundary detection of four-dimensional cardiac magnetic resonance image sequences". In: *IEEE transactions on medical imaging* 12.1, pp. 65–72 (cited on pp. 40, 41).

Sun, L. and Y. Wang (2018). "A multi-attribute fusion approach extending Dempster-Shafer theory for combinatorial-type evidences". In: *Expert Systems with Applications* 96, pp. 218–229 (cited on p. 23).

Sun, Yi, Chengfeng Zhou, Yanwei Fu, and Xiangyang Xue (2019). "Parasitic GAN for semi-supervised brain tumor segmentation". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. Taipei, Taiwan, pp. 1535–1539 (cited on pp. 61, 75).

Suykens, J.AK and J. Vandewalle (1999). "Least squares support vector machine classifiers". In: *Neural processing letters* 9.3, pp. 293–300 (cited on p. 13).

Taha, Abdel Aziz and Allan Hanbury (2015). "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC medical imaging* 15.1, pp. 1–28 (cited on p. 17).

Taleb-Ahmed, A. and L. Gautier (2002). "On information fusion to improve segmentation of MRI sequences". In: *Information Fusion* 3.2, pp. 103–117 (cited on pp. 45, 51).

Tavakoli, F. and J. Ghasemi (2018). "Brain MRI segmentation by combining different MRI modalities using Dempster-Shafer theory". In: *IET Image Processing* 12.8, pp. 1322–1330 (cited on pp. 45, 46).

Tong, Z. and T. Xu P.and Denœux (2021a). "An evidential classifier based on Dempster-Shafer theory and deep learning". In: *Neurocomputing* 450, pp. 275–293 (cited on p. 79).

Tong, Z., P. Xu, and T. Denœux (2021b). "Evidential fully convolutional network for semantic segmentation". In: *Applied Intelligence*, pp. 1–24 (cited on pp. 29, 79).

Trabelsi, O., L. Tlig, M. Sayadi, and F. Fnaiech (2015). "Skin lesion segmentation using the DS evidence theory based on the FCM using feature parameters". In: *12th International Multi-Conference on Systems, Signals & Devices (SSD15)*. IEEE. Sfax, Tunisia, pp. 1–5 (cited on pp. 45, 48).

Tran, Dustin, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner (2019). "Bayesian layers: A module for neural network uncertainty". In: *Advances in neural information processing systems* 32 (cited on p. 2).

Trebing, Kevin, Tomasz Stanczyk, and Siamak Mehrkanoon (2021). "SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture". In: *Pattern Recognition Letters* 145, pp. 178–186 (cited on p. 15).

Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky (2017). "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6924–6932 (cited on p. 83).

Vannoorenberghe, P., O. Colot, and D. De Brucq (1999). "Dempster-Shafer's theory as an aid to color information processing. Application to melanoma detection in dermatology". In: *Proceedings 10th International Conference on Image Analysis and Processing*. IEEE. Venice, Italy, pp. 774–779 (cited on pp. 45, 47).

Vannoorenberghe, P. and G. Flouzat (2006). "A belief-based pixel labeling strategy for medical and satellite image segmentation". In: *2006 IEEE International Conference on Fuzzy Systems*. IEEE. Vancouver, BC, Canada, pp. 1093–1098 (cited on p. 41).

Vauclin, S., P. Zhang, I. Gardin, O. O. Gallocher, and P. Vannoorenberghe (2005). "Segmentation of thoracic computed tomography images". In: *International Symposium on Signals, Circuits and Systems*. Vol. 1. Iasi, Romania, pp. 31–34. DOI: 10.1109/ISSCS.2005.1509843 (cited on p. 41).

Wang, Feifan, Runzhou Jiang, Liqin Zheng, Chun Meng, and Bharat Biswal (2019). "3d u-net based brain tumor segmentation and survival days prediction". In: *International MICCAI Brainlesion Workshop*. Springer. Shenzhen, China, pp. 131–141 (cited on p. 73).

Wang, R., X. Shen, Y. Li, Y. Zhu, C. Hui, and S. Zhang (2013). "Lesion segmentation in acute cerebral infarction based on Dempster-Shafer theory". In: *International Conference on Wavelet Analysis and Pattern Recognition*. IEEE. Zurich, Switzerland, pp. 209–214 (cited on p. 45).

Wen, J., Y. Tian, S. Yehang, T. Yongchuan, and H. Weiwei (2018). "Improved evidential fuzzy c-means method". In: *Journal of Systems Engineering and Electronics* 29.1, pp. 187–195 (cited on pp. 50, 53).

White, Franklin E (1991). *Data fusion lexicon*. Tech. rep. Joint Directors of Labs Washington DC (cited on p. 2).

Xi, Nan (2019). "Semi-supervised Attentive Mutual-info Generative Adversarial Network for Brain Tumor Segmentation". In: *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE. Wellington, New Zealand, pp. 1–7 (cited on p. 75).

Xiao, Fuyuan, Junhao Wen, and Witold Pedrycz (2022). "Generalized Divergence-based Decision Making Method with an Application to Pattern Classification". In: *IEEE Transactions on Knowledge and Data Engineering* (cited on p. 28).

Xiao, R., H. Ding, F. Zhai, T. Zhao, W. Zhou, and G. Wang (2017). "Vascular segmentation of head phase-contrast magnetic resonance angiograms using grayscale and shape features". In: *Computer Methods and Programs in Biomedicine* 142, pp. 157–166 (cited on pp. 45, 46).

Xin, G., Y. Xiao, and H. You (2005). "An improved Dempster-Shafer algorithm for resolving the conflicting evidences". In: *International Journal of Information Technology* 11.12, pp. 68–75 (cited on p. 52).

Xu, Ning, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang (2017). "Deep grabcut for object selection". In: *28th British Machine Vision Conference, BMVC 2017*. BMVA Press. London, UK (cited on p. 61).

Yager, R.-R (1987). "On the Dempster-Shafer framework and new combination rules". In: *Information sciences* 41.2, pp. 93–137 (cited on pp. 23, 24).

Yager, R.R (1992). "Decision making under Dempster-Shafer uncertainties". In: *International Journal of General System* 20.3, pp. 233–245 (cited on p. 28).

Yager, R.R. (2004). "Decision making using minimization of regret". In: *International Journal of Approximate Reasoning* 36.2, pp. 109–128. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2003.10.003. URL: https://www.sciencedirect.com/science/article/pii/S0888613X03001270 (cited on p. 28).

Yaniv, Ziv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare (2018). "SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research". In: *Journal of digital imaging* 31.3, pp. 290–303 (cited on p. 87).

Yu, Lequan, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng (2019). "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 605–613 (cited on p. 77).

Zaidi, Habib and Issam El Naqa (2010). "PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques". In: *European journal of nuclear medicine and molecular imaging* 37.11, pp. 2165–2187 (cited on p. 78).

Zeng, Guodong, Xin Yang, Jing Li, Lequan Yu, Pheng-Ann Heng, and Guoyan Zheng (2017). "3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images". In: *International workshop on machine learning in medical imaging*. Springer. Quebec City, Quebec, Canada, pp. 274–282 (cited on p. 75).

Zhou, Hong-Yu, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu (2021). "nnFormer: Interleaved Transformer for Volumetric Segmentation". In: *arXiv preprint arXiv:2109.03201* (cited on p. 112).

Zhou, T., S. Canu, and S. Ruan (2020). "Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation". In: *Computerized Medical Imaging and Graphics* 86, p. 101811 (cited on pp. 43, 105).

Zhou, Tongxue, Su Ruan, and Stéphane Canu (2019). "A review: Deep learning for medical image segmentation using multi-modality fusion". In: *Array* 3, p. 100004 (cited on p. 43).

Zhou, Tongxue, Su Ruan, Pierre Vera, and Stéphane Canu (2022). "A Tri-Attention fusion guided multi-modal segmentation network". In: *Pattern Recognition* 124, p. 108417 (cited on pp. 15, 43).

Zhu, W., Y. Huang, Z. Tang H.and Qian, W. Du N.and Fan, and X. Xie (2018). "Anatomynet: Deep 3D squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation". In: *bioRxiv*, p. 392969 (cited on p. 15).

Zhu, Y., L. Bentabet, O. Dupuis, V. Kaftandjian, D. Babot, and M. Rombaut (2002). "Automatic determination of mass functions in Dempster-Shafer theory using fuzzy C-means and spatial neighborhood information for image segmentation". In: *Optical Engineering* 41.4, pp. 760–770 (cited on pp. 31, 32, 45).

Zou, Ke, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu (2022). "TBraTS: Trusted Brain Tumor Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, pp. 503–513 (cited on p. 117).

Zouhal, L. Me. and T. Denœux (1998). "An evidence-theoretic k-NN rule with parameter optimization". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 28.2, pp. 263–271 (cited on p. 30).

List of Tables

List of Figures

List of Abbreviations

DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
BFT	Belief Function Theory
CT	Computed Tomography
PET	Positron Emission Tomography
MRI	Magnetic Resonance Imaging
FDG	Fluoro Deoxy Glucose
TE	Time Echo
RT	Repetition Time
FLAIR	FLuid-Attenuated Inversion Recovery
FDG	Ffluoro DeoxyGlucose
ROI	Region Of Interest
TP	True Positive
TF	True Negative
FP	False Positive
FN	False Negative
HD	Hausdorff Distance
ED	EDema
ET	Enhancing Tumor
NRC/NET	Necrotic Core/None Enhancing Tumor
BBA	Basic Belief Assignment
DM	Decision Making
EKNN	Evidential K-Nearest Neighbor
BFOD	Binary Frames Of Discernment
ENN	Evidential Neural Network

RBF	Radial Basis Function
FCM	Fuzzy C-Means
ECM	Evidential C-Means
GD	Gaussian Distribution
MV	Mebership Value
BraTs	Brain Tumor Segmentation
ECE	Expected Calibartion Error
MCD	Monte-Carlo Dropout
FE	Feature Extraction
SE	Evidential Segmentation
MMEF	MultiModal Evidence Fusion
CPU	Graphics Processor Unit
FLOPs	FLating point OPerations