UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE

HEUDIASYC (UTC) - LITIS(Univ-Rouen)

THÈSE

présentée en vue d'obtenir le titre de

Docteur de l'Université de Technologie de Compiègne

Spécialité : Technologies de l'Information et des Systèmes

Présentée par

CHUNFENG LIAN

Information Fusion and Decision-Making Using Belief Functions: Application to Therapeutic Monitoring of Cancer

Soutenue le 27 janvier 2017 devant le jury composé de :

M. Yves GRANDVALET	- DR, CNRS, Heudiasyc, UMR 7253	Président
M. Wojciech PIECZYNSKI	- PU, Télécom SudParis	Rapporteur
M. Olivier COLOT	- PU, Université Lille 1	Rapporteur
M. Rachid DERICHE	- DR, INRIA Sophia Antipolis	Examinateur
M. Christian BARILLOT	- DR, CNRS, IRISA, UMR 6074	Examinateur
Mme Su RUAN	- PU, Université de Rouen	Directrice
M. Thierry DENŒUX	- PU, UTC	Directeur

Information Fusion and Decision-Making Using Belief Functions: Application to Therapeutic Monitoring of Cancer

Thèse soutenue le 27 janvier 2017 devant le jury composé de :

M.	Yves GRANDVALET	DR, CNRS, Heudiasyc, UMR 7253	(Président du jury)
M.	Wojciech PIECZYNSKI	PU, Télécom SudParis	(Rapporteur)
M.	Olivier COLOT	PU, Université Lille 1	(Rapporteur)
M.	Rachid DERICHE	DR, INRIA Sophia Antipolis	(Examinateur)
M.	Christian BARILLOT	DR, CNRS, IRISA, UMR 6074	(Examinateur)
Mme	Su RUAN	PU, Université de Rouen	(Directrice de thèse)
M.	Thierry DENŒUX	PU, Université de Technologie de Compiègne	(Directeur de thèse)

To my grandmother, my parents, and my lovely wife

Abstract

Radiation therapy is one of the most principal options used in the treatment of malignant tumors. To enhance its effectiveness, two critical issues should be carefully dealt with, i.e., reliably predicting therapy outcomes to adapt undergoing treatment planning for individual patients, and accurately segmenting tumor volumes to maximize radiation delivery in tumor tissues while minimize side effects in adjacent organs at risk. Positron emission tomography with radioactive tracer fluorine-18 fluorodeoxyglucose (FDG-PET) can non-invasively provide significant information of the functional activities of tumor cells.

In this thesis, the goal of our study consists of two parts: 1) to propose reliable therapy outcome prediction system using primarily features extracted from FDG-PET images; 2) to propose automatic and accurate algorithms for tumor segmentation in PET and PET-CT images. The theory of belief functions is adopted in our study to model and reason with uncertain and imprecise knowledge quantified from noisy and blurring PET images. In the framework of belief functions, a sparse feature selection method and a low-rank metric learning method are proposed to improve the classification accuracy of the evidential K-nearest neighbor classifier learnt by high-dimensional data that contain unreliable features. Based on the above two theoretical studies, a robust prediction system is then proposed, in which the small-sized and imbalanced nature of clinical data is effectively tackled. To automatically delineate tumors in PET images, an unsupervised 3-D segmentation based on evidential clustering using the theory of belief functions and spatial information is proposed. This mono-modality segmentation method is then extended to co-segment tumor in PET-CT images, considering that these two distinct modalities contain complementary information to further improve the accuracy. All proposed methods have been performed on clinical data, giving better results comparing to the state of the art ones.

Keywords: Theory of belief functions, Feature selection, Distance metric learning, Data classification, Data clustering, Cancer therapy outcome prediction, Automatic tumor segmentation, PET/CT imaging

Résumé

La radiothérapie est une des méthodes principales utilisée dans le traitement thérapeutique des tumeurs malignes. Pour améliorer son efficacité, deux problèmes essentiels doivent être soigneusement traités : la prédication fiable des résultats thérapeutiques et la segmentation précise des volumes tumoraux. La tomographie d'émission de positrons au traceur Fluoro-18-déoxy-glucose (FDG-TEP) peut fournir de manière non invasive des informations significatives sur les activités fonctionnelles des cellules tumorales.

Les objectifs de cette thèse sont de proposer: 1) des systèmes fiables pour prédire les résultats du traitement contre le cancer en utilisant principalement des caractéristiques extraites des images FDG-TEP; 2) des algorithmes automatiques pour la segmentation de tumeurs de manière précise en TEP et TEP-TDM. La théorie des fonctions de croyance est choisie dans notre étude pour modéliser et raisonner des connaissances incertaines et imprécises pour des images TEP qui sont bruitées et floues. Dans le cadre des fonctions de croyance, nous proposons une méthode de sélection de caractéristiques de manière parcimonieuse et une méthode d'apprentissage de métriques permettant de rendre les classes bien séparées dans l'espace caractéristique afin d'améliorer la précision de classification du classificateur EK-NN. Basées sur ces deux études théoriques, un système robuste de prédiction est proposé, dans lequel le problème d'apprentissage pour des données de petite taille et déséquilibrées est traité de manière efficace. Pour segmenter automatiquement les tumeurs en TEP, une méthode 3-D non supervisée basée sur le regroupement évidentiel (evidential clustering) et l'information spatiale est proposée. Cette méthode de segmentation mono-modalité est ensuite étendue à la co-segmentation dans des images TEP-TDM, en considérant que ces deux modalités distinctes contiennent des informations complémentaires pour améliorer la précision. Toutes les méthodes proposées ont été testées sur des données cliniques, montrant leurs meilleures performances par rapport aux méthodes de l'état de l'art.

Mots-clés : La théorie de fonctions de croyance, Sélection des caractéristiques, Apprentissage de métriques, Classification des données, Clustering des données, Prédiction, Radiothérapie, Segmentation de tumeurs automatique, Imagerie TEP/TDM

Acknowledgements

First and foremost, I would like to thank my two excellent advisors, Prof. Thierry Denœux and Prof. Su Ruan. I am very grateful for having this precious opportunity to be guided by them, and be involved in this interesting PhD topic.

I would like to thank Prof. Ruan for her nuanced and constructive guidance. Thanks to her strong expertise and experience in medical image analysis, pattern recognition, and machine learning, I have been forcefully supervised to complete this thesis, and my research background has been enhanced to face future challenges. Prof. Ruan have taught me a lot of things, among which the most critical one that deeply influences me is her great responsibility and the quality of hardworking. I am also very grateful for her help and support in my private life during the last couple years. I do appreciate that she is always there when I am in trouble.

As a scrupulous and creative scholar working in the field of belief functions, pattern recognition, and machine learning, Prof. Denœux has taught me tremendous expertise, and he is always inspiring me to do better job by his insightful comments. The most remarkable thing that I have learnt from him is how to be a conscientious and professional researcher. I have learnt that I should pay attention to every detail during research, and only in this way could I find a solid way to move forward. The supervision of Prof. Denœux really means a lot to me, especially to my research career in the future.

I would also like to express my heartfelt gratitude for all the members of my thesis committee. I would like to thank Dr. Wojciech Pieczynski, Professor at Télécom ParisTech, and Dr. Olivier Colot, Professor at Université Lille 1, for spending their valuable time to review my thesis manuscript. I would also like to thank Dr. Christian Barillot, Director of Research at CNRS, UMR 6074 IRISA, Dr. Rachid Deriche, Director of Research at INRIA Sophia Antipolis, and Dr. Yves Grandvalet, Director of Research at CNRS, UMR 7253 Heudiasyc, for accepting to be examiners in my thesis defense.

I would like to say thank you to the doctors and medical physicists in the LITIS Quantif team that leading by Mr. Pierre Vera, Professor of University-Hospital Practitioner in

Centre Henri Becquerel. I am very grateful for their constructive comments on my work, and the support of their medical knowledge and clinical data.

I would also like to thank the China Scholarship Council for their financial support during my PhD study.

During the last couple of years, I have had a unforgettable wonderful time with my colleagues and friends in both the Heudiasyc Lab and LITIS Lab. I would like to thank them for their countless help, selfless support, warmheartedness, and encouragement.

Last but not the least, I would like to say thank you to my family, in particular to my grandmother, my parents and parents in law. Their unselfish love is my motivation to move forward. I would like to express my deepest gratitude to my beloved wife. She is my best friend and my source of strength.

List of Publications

International Journals

[1] Chunfeng Lian, Su Ruan, Thierry Denœux, Fabrice Jardin, and Pierre Vera, "Selecting Radiomic Features from FDG-PET Images for Cancer Treatment Outcome Prediction", *Medical Image Analysis*, Vol. 32, pages 257-268, 2016.

[2] Chunfeng Lian, Su Ruan, and Thierry Denœux, "Dissimilarity Metric Learning in the Belief Function Framework", *IEEE Transactions on Fuzzy Systems*, 2016 (in press).

[3] Chunfeng Lian, Su Ruan, and Thierry Denœux,"An Evidential Classifier based on Feature Selection and Two-Step Classification Strategy", *Pattern Recognition*, Vol. 48, pages 2318-2327, 2015.

[4] Chunfeng Lian, Su Ruan, Thierry Denœux, Hua Li, and Pierre Vera, "Spatial Evidential Clustering with Adaptive Distance Metric for Tumor Segmentation in FDG-PET Images", *IEEE Transactions on Biomedical Engineering*, (**under review**).

International Conferences

[1] Chunfeng Lian, Su Ruan, Thierry Denœux, Hua Li, and Pierre Vera, "Robust Cancer Treatment Outcome Prediction Dealing with Small-Sized and Imbalanced Data from FDG-PET Images", *19th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2016)*, Athens, Greece, October 2016.

[2] Chunfeng Lian, Su Ruan, Thierry Denœux, Hua Li, and Pierre Vera, "Dempster-Shafer Theory based Feature Selection with Sparse Constraint for Outcome Prediction in Cancer Therapy", *18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2015)*, Munich, Germany, October 2015.

[3] Chunfeng Lian, Su Ruan, Thierry Denœux, and Pierre Vera, "Outcome Prediction in Tumour Therapy based on Dempster-Shafer Theory", *IEEE 12th International Symposium on Biomedical Imaging (IEEE-ISBI 2015)*, New York, USA, April 2015.

[4] Chunfeng Lian, Su Ruan, and Thierry Denoeux, "Joint Feature Transformation and Selection based on Dempster-Shafer Theory", *16th International Conference on Information Processing and Management of Uncertainty In Knowledge-based Systems (IPMU 2016)*, Eindhoven, The Netherlands, June 2016.

[5] Chunfeng Lian, Su Ruan, Thierry Denœux, Hua Li, and Pierre Vera, "Tumor Delineation in FDG-PET Images Using A New Evidential Clustering Algorithm with Spatial Regularization And Adaptive Distance Metric", *IEEE 14th International Symposium on Biomedical Imaging (IEEE-ISBI 2017)*, Melbourne, Australia, April 2017 (accepted).

National Conference

- Chunfeng Lian, Su Ruan, Pierre Vera, and Thierry Denœux, "Dempster-Shafer Theory based Outcome Prediction in Cancer Therapy", Colloque Gretsi 2015, Lyon, France, September 2015.

Abstract submitted to International Medical Conference

- Chunfeng Lian et al., "Cancer Therapy Outcome Prediction based on Dempster-Shafer Theory and PET Imaging", *AAPM 57th Annual Meeting & Exhibition*, Anaheim California, July 2015. (*Won the finalist for the John R. Cameron young investigator competition.*)

Contents

List of Tables

List of Figures

List of Acronyms

ART	Adaptive radiation therapy
ANN	Artificial neural networks
AUC	Area Under the Curve
ADASYN	ADAptive SYNthetic sampling
BFT	Belief function theory
BK-NN	Belief-based K-Nearest Neighbor classifier
CT	Computed tomography
CCR	Credal classification rule
CTV	Clinical target volume
CART	Classification and regression tree
DSC	Dice's coefficient
DST	Dempster-Shafer theory
EK-NN	Evidential K-Nearest Neighbor classifier
EFS	Evidential feature selection
EDML	Evidential dissimilarity metric learning
EM	Expectation-maximization
ECM	Evidential C-means
ECM-MS	ECM integrating adaptive distance Metric and Spatial regularization
FAST	Feature Assessment by Sliding Thresholds
FBS	Forward-backward splitting
FCM	Fuzzy C-means
FCM-SW	FCM integrating Spatial information and à trous Wavelet transform
FDG	Fluorine-18(F-18) fluorodeoxyglucose
GTV	Gross tumor volume
GLCM	Gray-level co-occurrence matrix
GLSZM	Gray-level size-zone matrix
HD	Hausdorff distance

HFS	Hierarchical forward selection
IG	Information gain
IGRT	Image-guided radiation therapy
KCS	Kernel class separability
K-PCA	Kernel principal component analysis
LDA	Linear discriminant analysis
LOOCV	leave-one-out cross-validation
LMNN	Large margin nearest neighbor
MRF	Markov random field
MRI	Magnetic resonance imaging
MTV	Metabolic tumor volume
NCA	Neighborhood component analysis
NSCLC	Non-small cell lung cancer
OAR	Organs at risk
PCA	Principal component analysis
PTV	Planning target volume
PET	Positron emission tomography
RW	Random walks
RELIEF	RELevance In Estimating Features
ROC	Receiver operating characteristics
SECM	A Spatial version of ECM
SFS	Sequential forward selection
SBS	Sequential backward selection
SFFS	Sequential floating forward selection
SBR	Source-to-Background Ratio
SUV	Standard uptake values
SMOTE	Synthetic Minority Over-sampling TEchnique
SVMRFE	SVM integrating Recursive Feature Elimination
TAD	An adaptive thresholding method
TBM	Transferable belief model
TLG	Total lesion glycolysis
VOI	Volume of interest
3D-LARW	3D-locally adaptive random walk

Introduction

Context of the thesis

Cancer is a major public health problem over the world, especially in developing countries. According to regular investigation (normally every five years) by the International Agency for Research on Cancer (IARC), 8.2 million deaths worldwide in 2012 were due to cancer, where 65% cases were from developing countries [5]. As one of the most principal modalities used in the treatment of malignant tumors, radiation therapy (or radiotherapy) is received by almost 50% of all cancer patients, and lead to 40% of curative treatment [6]. Due to sustained advancement of medical imaging techniques, as well as progresses made in understanding the radiobiology, the effectiveness of radiation therapy is being increasing enhanced.

Positron emission tomography (PET) is an advanced functional and molecular imaging tool generally used in cancer diagnosis, staging, and radiation oncology. As it can monitor functional activities of tumor cells *in vivo*, PET scanning with radioactive tracer fluorine-18 (F-18) fluorodeoxyglucose (FDG), i.e., FDG-PET, is playing a significant role in multiple tasks of radiation therapy, including reliable prediction of therapy outcomes to adapt treatment planning for individual patient, and accurate segmentation of target tumor volumes to maximize treatment effectiveness while minimize side effects in organs at risk.

Reliably predicting treatment outcomes before or even during radiation therapy is of great clinical value, as it can provide critical evidence to help re-optimizing the initial treatment plan for individual patient. While the analysis of acquired PET images has been claimed to be useful in this task, the solid application of PET-based therapy outcome prediction is hampered by some practical challenges, including imprecise and unreliable image features caused by noise and blur of PET imaging system, and small-sized and imbalanced datasets that can be gathered for training a well-performed prediction model.

Accurately delineating tumor volumes in PET images is beneficial for effective radiation therapy, as PET images can present precise information regarding heterogeneous biological

activities of tumor cells, thus providing specific knowledge to help defining inhomogeneous delivery of radiation dose. However, the automatic and accurate tumor segmentation in PET images is still an open issue, as PET images are blurring and noisy. Moreover, recent advancements in supervised learning community, especially in deep learning, are not applicable to this task, since tumors are inhomogeneous with arbitrary shapes, and they are varying from one patient to another.

The goal of our study in this thesis is thus to 1) develop reliable models for radiation therapy outcome prediction using primarily radiomic features extracted from FDG-PET images, and 2) propose automatic segmentation algorithms in 3-D for accurate delineation of tumor volumes in PET and PET-CT images. The basis of our study is the theory of belief functions, which is a powerful framework for modeling, fusing and reasoning with uncertain and/or imprecise information, such as that presented in PET images.

Contributions of the thesis

Theoretical studies

- *Supervised learning approaches:* A new fusion method, called "Dempster+Yager" mixed combination rule, for robustly representing uncertainty and imprecision of studied datasets in the evidential K-nearest neighbor (EK-NN) classification rule; An evidential feature selection (EFS) method with sparsity constraint for selecting informative feature subsets to reduce data imprecision and improve classification performance of the EK-NN method; An evidential dissimilarity metric learning (EDML) method with specific sparsity regularization for low-dimensional feature transformation and feature selection, so as to maximize the accuracy and efficiency of the EK-NN classification.

- *Unsupervised learning methods:* An evidential clustering algorithm integrating adaptive distance measure, feature selection, and MRF-based spatial regularization for image segmentation; An extension of this evidential clustering algorithm with specific consistency quantification and iterative information fusion for multi-modality segmentation.

Applications

- New prior knowledge definition and data rebalancing procedure to deal with small-sized and imbalanced learning problem in developing prediction models. Cancer therapy outcome prediction based on the above supervised learning approaches for treatment planning adaptation in radiation therapy.

- Automatic tumor segmentation in PET images and co-segmentation in PET-CT images using the above unsupervised learning algorithms for target tumor definition in radiation therapy.

Layout of the thesis

This thesis is structured in three parts, covering seven chapters:

- Part I presents the general context of radiation therapy, and the fundamental background of belief function theory. Chapter 1 describes the principles of radiation therapy, the clinical roles of PET/CT imaging in radiation therapy, and the challenges for therapy outcome prediction using PET images and automatic target tumor delineation in PET images. Chapter 2 introduces the main components of belief function theory, and the applications of belief functions in data classification and clustering.

- Part II introduces three methods for cancer therapy outcome prediction using belief functions and PET images. Chapter 3 proposes an evidential feature selection (EFS) method to improve classification accuracy of a variant evidential K-nearest neighbor (EK-NN) classifier. Chapter 4 proposes an evidential dissimilarity metric learning (EDML) method to maximize the performance of the EK-NN method, which attempts to learn a low-dimensional feature transformation from original feature space. Chapter 5 proposes solutions to deal with imbalanced and small-sized nature of studied clinical datasets, so as to select reliable feature subsets for cancer treatment outcome prediction.

- Part III deals with automatic tumor segmentation in PET and PET-CT images using belief functions. Chapter 6 presents an automatic segmentation method based on evidential clustering for segmenting tumor in PET images. This method is then extended in Chapter 7 to realize joint tumor segmentation in PET-CT images, so as

to combine complementary information from the two distinct image modalities for more accurate target tumor definition.

It is worth indicating that Chapter 3 to Chapter 6 are excerpted from four different papers that have been published or submitted.

Finally, we conclude our work in this thesis, and present the perspectives for our work.

Part I

General Context and Theoretical Background

Radiation Therapy and Challenges

Radiation therapy, also known as radiotherapy, is one of the five principal modalities used in the treatment of malignant tumors, the other four being surgery, chemotherapy, immunotherapy, and hormonal therapy. Almost 50% of all cancer patients received radiation therapy, leading to 40% of curative treatment for cancer [6]. Thanks to sustained advancement of medical imaging techniques, as well as progresses made in understanding the radiobiology, the effectiveness of radiation therapy for cancer treatment is still being continuously improved.

1.1 Introduction of Radiation Therapy

1.1.1 Principle of Radiation Therapy

Radiation therapy can be performed as the intent of cure, or can also be used as a complementary for other treatment modalities (e.g. surgery and chemotherapy). It utilizes ionizing radiation, a physical agent, to destroy the multiplication of cancer cells. By depositing high energy in the cells of the tissues where it passes through, radiation can kill cancer cells directly or lead to genetic changes in cancer cells to block them from further proliferating. While high-energy radiation damages both normal cells and tumor cells, due to the fact that tumor cells are slower than normal ones in repairing the functional damage, they are more likely to be killed by radiation therapy [6]. However, it is still necessary to avoid additional damages to organs at risk as less as possible.

External beam radiation is the most commonly used approach in nowadays clinical setting of radiation therapy. In external beam radiation therapy, the radiation source is outside the human body, and the target within the body is irradiated with external radiation beams (e.g. photon beams, and electron beams, etc). The other approach of radiation therapy is brachytherapy, which directly places radiation material into or close

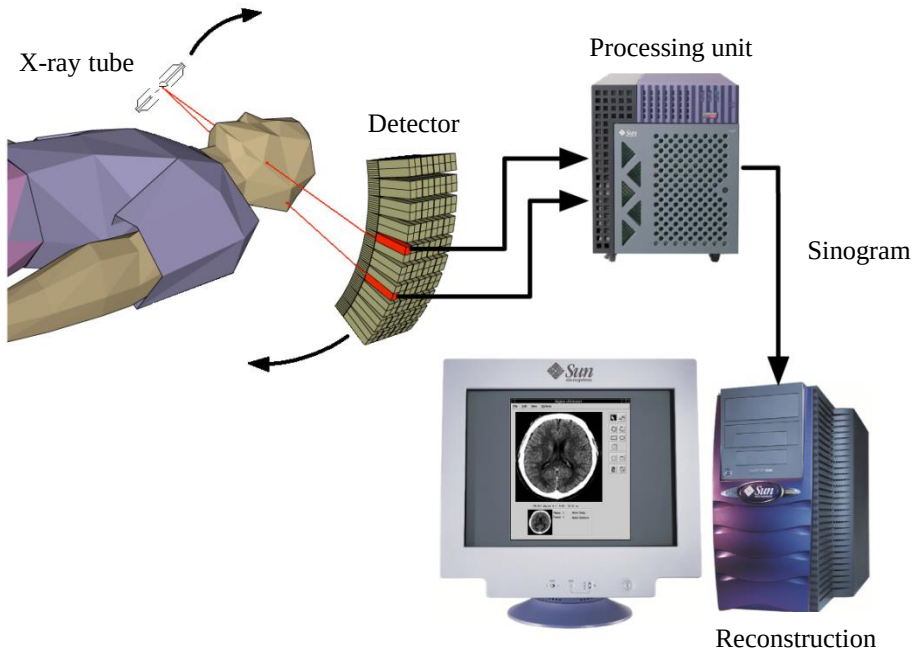


Figure 1.1: Principle of CT imaging. The X-ray generator emit X photons. These photons are attenuated when passing through the human body, and finally catched by the detectors. The X-ray generator and the detectors are rotating together to obtain signals from different angles.

to the target volume [7].

1.1.2 Role of PET/CT Imaging in Radiation Therapy

Medical imaging is playing an increasing central role in radiation therapy practice, due to advances in imaging techniques as well as progresses in radiation oncology. It actually influences the effectiveness of almost all available aspects of a radiation therapy protocol [8], including delineation of radiation target volumes and adjacent normal tissues, design of the radiation dose distribution in the planning process, and monitoring of treatment response during radiation therapy, etc. An approach that adopts advanced imaging technology to reduce uncertainties and assist decision making during a course of treatment is often referred to as image-guided radiation therapy (IGRT) [9].

As a device widely available at almost all cancer centers, computed tomography (CT) is the gold standard image modality in radiation oncology [10]. CT images can provide anatomical information to show geometric positions of target tumor and adjacent organs at risk. The principle of CT imaging can be briefly described in Figure 1.1. The use of

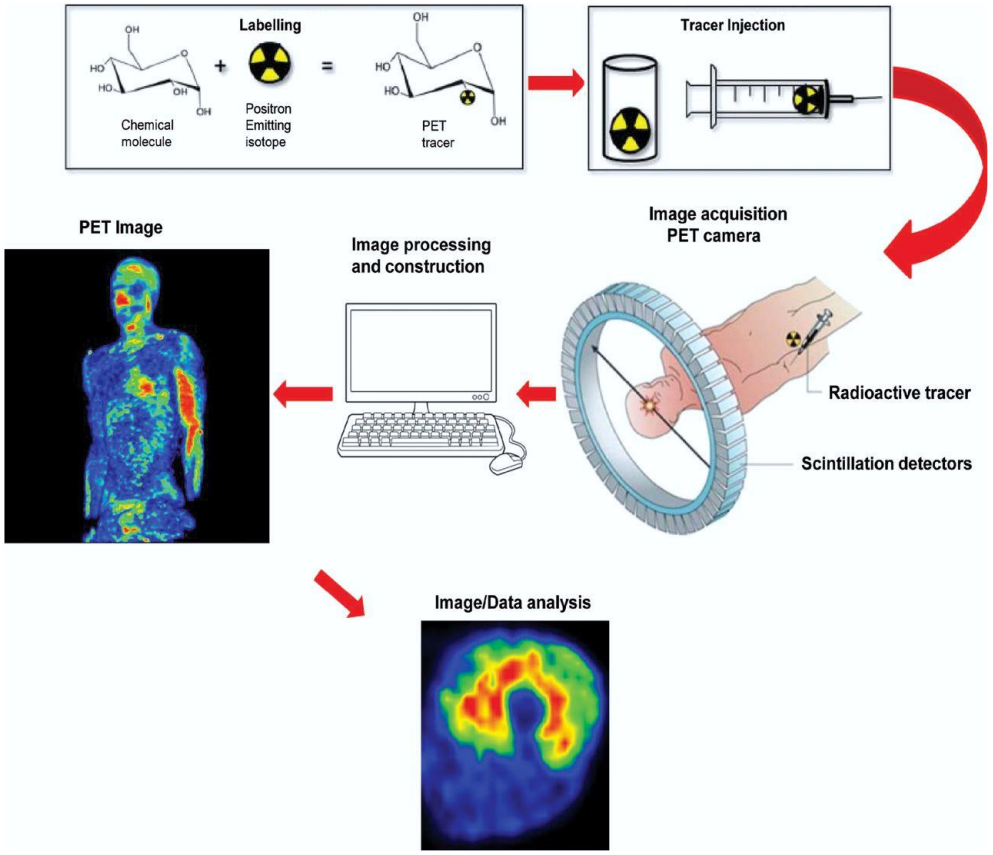


Figure 1.2: Principle of PET imaging [4]. The atoms of injected radioactive tracer emit positrons. The annihilation coincidences caused by collisions between positrons and electrons are then detected by scintillation detectors of a PET scanner, which are used to reconstruct images of the original distribution of injected biologically active molecule.

CT in radiation therapy allows three-dimensional dose calculation, dose optimization, and patient positioning [11]. Complementary to CT, positron emission tomography (PET) can provide critical functional information of target tumor for more precisely guarding the procedure of radiation therapy.

PET is a functional imaging technique used in nuclear medicine that can measure tissue metabolic activity *in vivo* through an injected radioactive tracer. The principle of PET imaging can be briefly described in Figure 1.2. With different radioactive tracers, PET imaging is endowed the ability to monitor different functional activity of a target tumor (e.g. metabolism, proliferation, and oxygen delivery) in molecular scale. In clinical oncology practice, the most commonly used radioactive tracer is fluorine-18 (F-18) fluorodeoxyglucose (FDG). PET scanning with FDG, i.e., FDG-PET, can highlight tumor tissues with high metabolic rate, thus has been widely used for diagnosis, staging, and re-staging

of most cancers, such as non-small cell lung cancer [12], esophageal carcinoma [13], or lymphomas [14], etc.

Apart from the above applications, FDG-PET imaging is also playing a significant role in radiation therapy:

- FDG-PET can provide molecular information, as a complement to anatomical information offered by CT, to assist the definition of radiation target volume in treatment planning process [15]. It has been proven that the combination of PET and CT can effectively reduce inter- and intra-operator variability in tumor delineation [16].

- FDG-PET can be adopted to identify a high-uptake subvolume within a gross tumor delineated in CT images, which will be irradiated by a higher dose of radiation. This procedure is often referred to as dose painting by contours or subvolume boosting [17], which allows for an inhomogeneous delivery of radiation dose, thus taking into account the heterogenous activity of tumor cells in radiation therapy.

- FDG-PET can also be used in the follow-up and evaluation of radiation therapy outcomes [18]. Considering that the metabolism changes of a tumor are usually prior to the morphological modifications, FDG-PET may provide more earlier detection of tumor response to the undergoing therapy than CT.

- Furthermore, increasing studies, e.g., [19–23], have shown that the functional information provided by FDG-PET images can predict treatment outcomes before the accomplishment of radiation therapy, offering promising evidence for the adaptation of a more effective treatment plan for individual patient.

It is thus of great clinical value to integrate FDG-PET imaging into radiation therapy in terms of the definition and adaptation of treatment planning, and also of the evaluation and prediction of treatment outcomes.

1.2 Treatment Planning of Radiation Therapy

Before starting treatment, the delivery of external radiation beam should be carefully planned. At this stage, doctors hope to find an ideal treatment position and the exact area to be irradiated for the patient, aiming to ensure that the tumor gets the prescribed dose of radiation while the surrounding normal tissues get as little as possible. Three critical issues

should be tackled, namely the accurate delineation of target tumor volumes and organs at risk surrounding the target, the complete prescription of radiation dose, and the effective adaptation of undergoing treatment plan for the patient.

1.2.1 Definition of Standardized Uptake Values in PET

FDG-PET imaging provides promising metabolic information of tumor cells for target tumor definition, does prescription, and treatment planning adaptation. The quantification of FDG uptake in PET images is usually performed in terms of standardized uptake values (SUV), as it offers a physiologically relevant measurement of cellular metabolism. SUV denotes the activity concentration of FDG whithin a lesion, normalized by the decay-corrected injected dose per unit body volume. Practically, patient body volume is usually surrogated by body weight, or body surface area. SUV normalized to body weight is given by

$$\text{SUV} = \frac{\text{Radioactivity concentration per unit volume (MBq/mL)}}{\text{Injected dose (MBq)/Body weight (g)}}, \tag{1.1}$$

where the unit of SUV is g/mL. Ideally, the utilization of SUV can remove quantification variability that caused by differences in patient size and the amount of injected dose.

1.2.2 Definition of Target Volumes

The size, shape, and location of target tumor volumes and surrounding organs at risk are usually defined in the 3-D model constructed by the planning CT images. The accurate delineation requires comprehensively taking into account all available knowledge regarding a tumor, which includes the anatomical information provided by CT images, the soft tissue composition information revealed by magnetic resonance (MR) images, and the functional activities of tumor tissues reflected by PET images, etc..

In the process of delineating the target tumor, different volumes should be defined due to various reasons, e.g., probable movement of patient during treatment, varying concentrations of malignant cells, and potential change of the spatial relationship between tumor volume and radiation beam during treatment, etc.. According to two reports presented by the International Commission on Radiation Units and Measurements, i.e., ICRU Reports No. 50 [24] and No. 62 [25], several principal and critical structure volumes are defined, which include gross tumor volume (GTV), clinical target volume (CTV), planning target volume (PTV), and organs at risk (OARs), so as to aid in the treatment planning process,

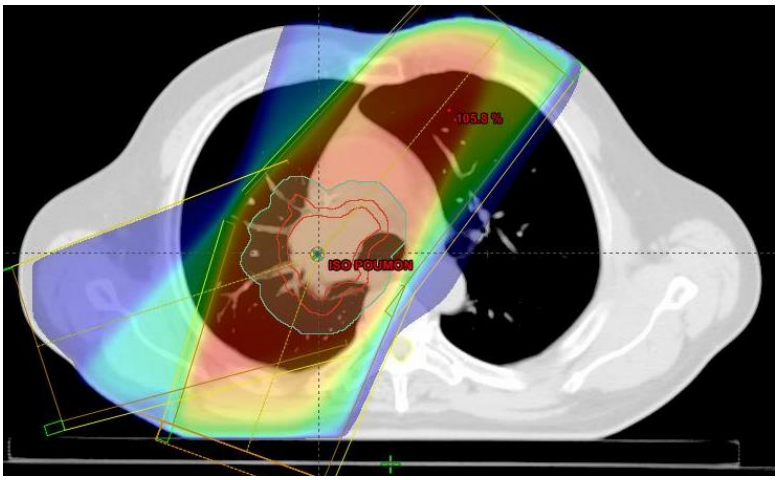


Figure 1.3: Definition of GTV (inside boundary), CTV (intermediate boundary), PTV (external boundary), and OARs (normal lung tissues adjacent to PTV) in an axial CT slice for a lung tumor example.

and to provide a basis for comparison of treatment outcomes [7]. An example of these critical volumes are illustrated in Figure 1.3. The definition of them can be described as follows:

- The gross tumor volume (GTV) is the gross palpable or visible extent and location of malignant growth. The definition of GTV is usually based on multi-sources of information, including multi-modality medical images, clinical examination, and diagnostic knowledge (e.g. histological reports). Abounding research, e.g., [15,16,26–29], has shown that integrating functional knowledge provided by PET with the anatomical information offered by CT can effectively improve the reliability of delineated GTV.

- The clinical target volume (CTV) is usually defined by adding an empirical margin concentrically to the GTV. Apart from the contained demonstrable GTV, it may also include sub-clinical microscopic extension of the primary tumor or regional lymph node spread.

- The planning target volume (PTV) includes CTV and an additional margin accounting for intra-treatment variations, inherent uncertainties in therapy setup, potential organ motions, and machine tolerances.

- The organs at risk (OARs) are normal tissues that close to the PTV. As OARs are usually sensitive to radiation, specific efforts should be paid to minimize irradiation

of them to avoid substantial morbidity.

In general, the accurate delineation of target tumor volumes and OARs plays a central role in the process of radiation therapy planning. The aim should be to deliver a sufficient high dose of radiation to the tumor, while as low as possible a dose to the OARs.

1.2.3 Adaptation of Treatment Plan

The traditional treatment plan is usually fixed under the assumption that the safety margin added to the CTV/GTV can properly handle positioning uncertainties and biological changes of patients during the entire course of radiation therapy. However, as the margin is defined based on the standard deviation of positional variation averaged from patient populations, it ignores the specificity of each individual patient, such as the variations in biological and morphological change, or the dosimetric variation in organs and targets of interest [30]. Thus, an adaptable treatment plan taking into account these specific changes of the individual patient should be more appropriate for effective radiation therapy.

To improve treatment outcomes, the concept of adaptive radiation therapy (ART) was first introduced by Yan et al. in [31], where treatment variations were systematically monitored to re-optimize the undergoing plan early on during the course of treatment. Customizing treatment dose and target margin for each individual patient is of great value for adaptive radiation therapy [32]. Usually, by integrating of advanced technologies (e.g. PET-CT scanners) in radiation oncology clinic, the dose of radiation can be modified spatially and/or temporally during the treatment process.

The longitudinal modification of a undergoing treatment plan depends heavily on the reliable monitor and prediction of tumor response to treatment for the individual patient. Responses of tumors to an identical therapy vary among patients. As an advanced imaging tool that can sensitively monitor pathologic response of tumor cells for the delivered radiation, PET imaging has a significant impact on updating dose distribution and delineating target volumes in adaptive radiation therapy, such as the example presented in [33].

To sum up, based on the inclusion of additional instruments, e.g., the integration of function PET imaging with anatomical CT imaging, adaptive radiation therapy can realize the modification and adjustment of an initial treatment plan early on during the treatment process, thus potentially improving outcomes of radiation therapy.

1.3 Assessment and Follow-Up of Treatment Outcomes

Radiation therapy outcomes can be defined as tumour response to delivered radiation, toxicity evolution during follow-up, rates of local recurrence, evolution to metastatic disease, survival or a combination of these factors [34]. The definition of treatment outcomes for solid tumors is based on assessment of target tumor burden and its change in study [35]. Baseline evaluations of tumor lesions should be performed closely to the beginning of radiation therapy, with a gap of time being shorter than 4 weeks. Then, to follow and monitor changes of target tumor for on-going treatment, evaluations should be kept the same as that in the baseline.

In clinical practice, imaging based evaluation is always preferable than clinical examination, unless the target tumor being monitored becomes non-visible in medical images but is measurable by clinical exam. The advantage of evaluation using medical images is non-invasive and objective; moreover, acquired medical images can be rechecked after current evaluation. FDG-PET/CT scanner has specific advantages in outcome evaluation [36]. For example, post-therapy changes of target tumor, e.g., fibrosis and necrosis that obscure the identification of recurrent tumor in CT, could be sensitively characterized with FDG-PET. The metabolic change of target tumor captured by FDG-PET may be a better indicator of a favorable response to therapy than size change of target tumor reflected by CT [37].

1.4 Prediction of Treatment Outcomes

Accurately predicting treatment outcomes plays a significant role for the development of individualized medicine [34]. For instance, if the outcomes after a treatment for a specific patient can be reliably forecasted before or during the process of radiation therapy, clinicians can then update and re-optimize on-going treatment plan for this individual patient. To this end, diversity and heterogenous medical data of cancer patients can be gathered to develop a prediction model.

1.4.1 Radiation Therapy Outcome Prediction in Clinical Study

With increasing advancements of techniques in medical imaging, in image-guided radiation therapy (IGRT), as well as in image-guided adaptive radiation therapy (IGART), medical image-based factors, or noninvasive imaging biomarkers, tend to be more and more reliable cancer treatment outcome predictors in clinical study.

As an advanced imaging tool generally used in clinical oncology, FDG-PET imaging has been proven by many researchers to be predictive of pathologic response of tumor cells for undergoing treatment. The pretreatment maximum standard uptake value (SUV_{max}) is strongly relevant to tumor recurrence and survival in lung cancers [38], cervical cancers [39], and head and neck cancers [40], etc.. In [41], Lemarignier et al. have proven that metabolic tumor volume (MTV) and total lesion glycolysis (TLG), two pretherapy quantitative metabolic parameters, correlate with treatment outcomes in patients with oesophageal squamous cell carcinoma. In [42], Vera et al. have validated the predictive value of SUV_{max} assessed during the 5th week of curative-intent radiation therapy in patients with non-small-cell lung cancer. Some other studies have also shown that longitudinal changes of SUV between different time points are early predictors of therapy outcomes [43, 44].

1.4.2 Radiomics-Based Treatment Outcome Prediction

Radiomics refers to the analysis and mining of high-dimensional quantitative features extracted from medical images. It provides an unprecedented opportunity to support and improve personalized clinical decision making [45, 46]. Building upon advanced machine learning and pattern recognition techniques, the ultimate goal of radiomics analysis is to incorporate critical imaging information into prediction models, so as to provide added value for commonly used predictors (e.g. genomics analysis), and to improve the prediction of treatment outcomes [47].

In radiomics analysis, the hypothesize is that a high-dimensional and minable feature space obtained by automatic imaging feature extraction algorithms can capture inter- and intra- tumor heterogeneity, and thus may have great predictive power [48]. Plenty of studies have been performed to evaluate the correlation between treatment outcomes and first-, second-, and higher-order imaging features extracted from FDG-PET images.

For instance, to predict treatment outcomes for patients with cervix cancer or head and neck cancer, El Naqa et al. [19] have investigated the predictive power of logistic regression model [49] trained by shape and textural features extracted from PET images; In [50] and [21], temporal changes of FDG-PET features have been adopted to predict pathologic response of esophageal cancer to chemoradiation therapy using support vector machine [51] or logistic regression model; In [20], Tixier et al. have attempted to characterize intra-tumor inhomogeneity and predict response to radiochemotherapy in esophageal cancer based on textural features extracted from baseline FDG-PET images; while, in [23], a

logistic regression model based on joint quantification of FDG-PET and MRI textures has been developed for the prediction of lung metastases in soft-tissue sarcomas of the extremities.

All the studies mentioned above demonstrate the significant role of radiomics analysis in predicting treatment outcomes of radiation therapy.

1.4.3 Challenges for Reliable Prediction of Treatment Outcomes

While the quantification of PET image features in radiomics analysis has beem claimed to be useful in cancer treatment outcome prediction, its solid application is still hampered by multiple practical challenges caused by low-quality of PET images, variances of tumor sizes and shapes, and limited number of observations, etc.

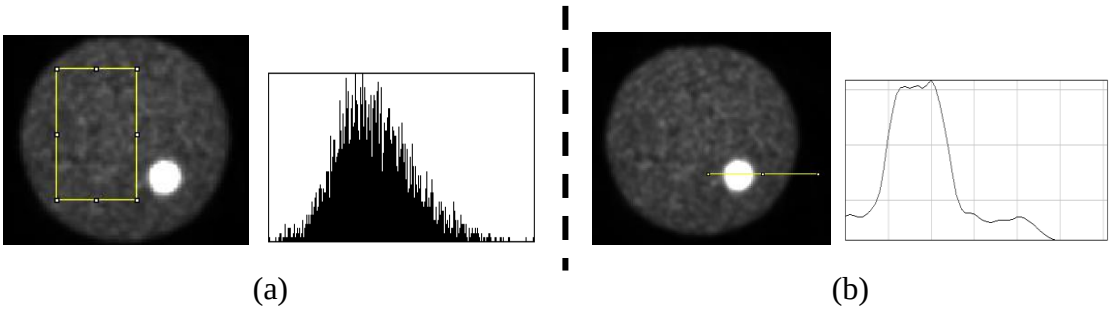


Figure 1.4: An example of phantom PET image, where (a) and (b) present the histograms of two different regions to show, respectively, the uncertainty and imprecision nature of PET imaging.

- *Uncertainty and imprecision of extracted features:* PET images are noisy and blurring, which are mainly caused by inaccurate image acquisition system and limited spatial resolution, respectively [52]. As an example shown in Figure 1.4, due to noise and blur in PET imaging system, reconstructed image often contains uncertain and imprecise information. As the result, some features extracted from PET images may be unreliable for accurate treatment outcome prediction.

  In addition, a large amount of features can be extracted from multi-sources of information (e.g. multi-modality medical images), while there is no consensus regarding the most informative ones. Some of these features may be redundant or irrelevant to treatment outcomes. A worse thing is that badly defined features may even degrade prediction accuracy when they were included in a prediction model [53].

The effect of small tumor volumes [54] may also lead to unreliable texture analysis and bias quantification of intra-tumoral heterogeneity, thus decreasing the performance of a prediction model.

- *Effect of small-sized and imbalanced data:* Comparing to a relatively high-dimensional feature space, only a limited number of observations (or training samples) are available for constructing a cancer treatment outcome prediction model. As a challenge often encountered in the medical domain, small sized learning set may hamper the performance of traditional classification models, since these classifiers had been proposed under the assumption that adequate training instances can be gathered for learning. In addition, a high dimensional feature space may further increase the complexity of these learning models, thus leading to high risk of over-fitting on small-sized training set [55].

  Imbalanced or skewed dataset, where different classes have distinct number of training samples, is also a typical problem of medical data. Since most standard learning algorithms had been designed for well-balanced datasets, the complex class imbalance challenge could hinder them from properly representing the distributive characteristics of the data [56], thus leading to unfavorable prediction accuracies across patients with different treatment outcomes. For example, assume we have a learning set consisting of 50 lung tumor patients, where 45 patients suffered from tumor recurrence after treatment, while the other 5 patients were no-recurrence. Using average classification accuracy as the criterion, a traditional classification model trained on this dataset may lead to "high" prediction accuracy (90%), but at the cost of improperly assigning all patients into the majority class (i.e. recurrence).

Therefore, in order to construct a promising prediction model, it is required to select the most informative features from uncertain and imprecise input space; in addition, the influence of imbalanced and small-sized training data on robust feature selection should also be effectively tackled.

1.5 Automatic Delineation of Tumor Volumes

Accurate tumor segmentation in PET images is a vital step for diverse objectives in clinical oncology, including reliable diagnosis and tumor staging, as well as solid radiation therapy

treatment planning. Manually delineating contours around a target is the most intuitive and common way in clinical practice. However, since manual delineation suffers from high intra- and inter-operator variability, developing automatic or semi-automatic methods is necessary for objective tumor segmentation.

1.5.1 Significance of Automatic Segmentation for Radiation Therapy

In the practical process of radiation therapy planning, the definition of target tumor volumes is usually carried out manually by experienced clinicians with a computer interface. Due to the blurring nature of PET images that caused by low spatial resolution and partial volume effect, manual segmentation is highly subjective and suffers from many drawbacks, including time consuming, labor intensive, and operator-dependent [57].

The severe inter- and intra-operator variability of manual segmentation may lead to imprecise and unreproducible tumor contours. For instance, in two different studies of lung lesion delineation, i.e., [58] and [59], both high intra- and inter-operator variability were observed, where the former variability was $42-84\%$ and $0-44\%$, respectively; while, the latter variability reached $44-78\%$ and $0-66\%$, respectively.

The problem of manual segmentation emphasizes the requirement of objective and reliable automatic or semi-automatic algorithms.

1.5.2 Automatic Algorithms for Tumor Segmentation

Diverse automatic or semi-automatic PET image segmentation algorithms have been proposed, such as thresholding methods, region growing methods, statistical methods, graph-based methods, or clustering methods, etc. Although supervised learning methods, especially deep learning models, have been successfully used in multiple research domain, they are not applicable in automatic tumor segmentation, since tumors present heterogenous uptakes and irregular boundaries.

Thresholding methods usually define a fixed, an adaptive, or an iterative threshold to differentiate lesions from background. In fixed thresholding, a constant threshold value needs to be determined in terms of standard uptake value (SUV). Typically, a threshold as 40% of the maximum SUV (SUV_{max}) is adopted to segment positive tissues for lung cancer, head and neck cancer, and cervical cancer [60]. The adaptive thresholding methods improve fixed ones by selecting threshold values according to specific criteria with respect to, e.g., source-to-background ratio (SBR) [61], mean target SUV [62], and scanner resolution [63].

The iterative thresholding methods [64] usually determine the optimum threshold for PET image segmentation based on calibrated threshold-volume curves acquired by phantom studies. While thresholding methods are simple and intuitive, they are sensitive to noise, and incapable to handle intensity variations, making them challenged by irregularly shaped heterogeneous uptake distributions [57].

Started by the initialization of seed points, region growing methods [65, 66] delineate tumor by repeatedly including or excluding adjacent voxels according to predefined criteria. The definition of a criterion for region growing is often based on spatial context, thus usually working well in segmenting homogenous targets. However, the performance of these methods depends heavily on the quality of initialization, and may fail to segment heterogeneous objects.

Statistical methods delineate tumors according to the assumption that target and background obey distinct statistical distributions. For instance, Aristophanous et al. [67] regarded intensities of PET voxels as observations generated by a mixture of Gaussians. The parameters of each Gaussian was then calculated by the expectation-maximization (EM) algorithm to output segmentation results. While statistical methods are robust to noise and partial volume effect caused by low-resolution imaging system, they are sensitive to heterogeneous uptake of positive tissues.

Graph-based segmentation methods regard a given PET image as a graph constructed by different nodes and edges, where nodes are image voxels, while each edge usually quantifies the dissimilarity between two different voxels. Graph-cut [68] and random walk [69] are two important graph-based methods, and both of them have been applied to segmentation tumors in PET or PET-CT images [26–29, 70, 71]. While they can effectively combine global cue with local smoothness, their performance is strongly influenced by the quality of foreground and background seeds.

Unlike supervised learning methods that need a training step, clustering methods are suitable for PET image segmentation, because the positive tissues are inhomogeneous with non-convex shapes and vary according to patients [57]. In view of the wide applications of fuzzy c-means (FCM) in medical image segmentation tasks [72–74], Belhassen et al. have proposed a robust approach, called FCM-SW [75], working specifically for the segmentation of heterogeneous tumors in PET images. As an extension of FCM and possibilistic clustering [76], an evidential c-means algorithm (ECM), based on the theory of belief functions (BFT) [77], has been proposed in [78], and has been extended to segment multi-parameter

MR images [79]. A spatial version of ECM, namely SECM [80], has then been developed recently for lung tumor delineation in multi-tracer PET images. One of the key issues to ensure the performance of clustering methods is to define a robust distance metric which can reliably quantify clustering distortions and local smoothness.

1.5.3 Challenges for Accurate Segmentation of Tumor Volumes

Although diverse automatic or semi-automatic algorithms have been proposed, the accurate segmentation of tumor volumes in PET images is still an open issue. To further improve the reliability of target tumors delineated in PET images, some critical challenges should be carefully dealt with:

- Reliably modeling of uncertainty and imprecision inherent in PET, since PET images are noisy and blurring due to inaccurate imaging acquisition system and partial volume effect caused by low spatial resolution.
- Effectively quantifying context information, because tumors shown in PET images usually present heterogenous distribution of radioactivity. This challenge strongly hampers the performance of available automatic methods that using only intensity values for target delineation.
- Taking into account other image modalities (e.g. CT images produced by integrated PET-CT scanner) that may provide additional knowledge to improve tumor delineation in PET images. The challenge is to find an appropriate way to fuse them with PET, since these distinct information sources concerning the same target tumor may be not always complementary, while sometimes may also partially contradicts with each other.

1.6 Propositions

As shown in Section 1.4 and Section 1.5, both PET imaging based treatment outcome prediction and automatic tumor segmentation in PET images play significant roles in radiation therapy. To ensure the effectiveness of radiation therapy, the specific challenges have been discussed in Section 1.4.3 and Section 1.5.3, respectively. The goal of our study in this thesis is thus to develop robust prediction models and reliable tumor segmentation algorithms. To these ends, our propositions can be briefly summarized as follows.

First of all, effective modeling of uncertainty and imprecision is a precondition for both two tasks. As the theory of belief functions (BFT) [77,81] is a formal and powerful tool for modeling, fusing, and reasoning with uncertain and/or imprecise information, we choose it to tackle this critical issue.

1.6.1 Propositions for Reliable Cancer Treatment Outcome Prediction

Dealing with low-quality original data that contain unreliable and imprecise features extracted from multi-sources of information, we propose a feature selection method and a dissimilarity metric learning method, both based on belief functions, so as to maximize (minimize) the impact of informative features (unreliable features) on constructing stable classification models for cancer treatment outcome prediction. The performance of the proposed feature selection method on small-sized and imbalanced clinical datasets then be further improved by including specific prior knowledge and data rebalancing procedure.

1.6.2 Propositions for Automatic Tumor Segmentation in PET Images

Considering clustering algorithms are suitable for PET image segmentation, especially when the positive tissues are inhomogeneous with non-convex shapes, an automatic segmentation method based on clustering is developed in 3-D, where, different from available methods, PET voxels are described not only by intensities but also complementally by image features. A specific procedure is adopted to select the most informative image features for voxel clustering, and to adapt distance metric for reliably representing clustering distortions and neighborhood similarities. A specific spatial regularization is also included in the clustering algorithm to effectively quantify local homogeneity.

This segmentation algorithm then be extended to jointly segment tumor in PET-CT images. By iteratively consistence quantification and information fusion in the co-segmentation framework, complementary knowledge in CT is effectively combined with that in PET to further improve the segmentation performance.

1.7 Conclusion

In this chapter, the principles of radiation therapy, and the significant role of FDG-PET imaging in radiation therapy have been briefly described. As two important tasks

to improve the effectiveness of radiation therapy, prediction of treatment outcomes, and automatic segmentation of tumor volumes have then been introduced.

The goal of our study is to develop robust outcome prediction models and automatic tumor segmentation algorithms, so as to improve the effectiveness of radiation therapy for individual patient. All the methods that will be introduced in the sequel are developed in the framework of belief functions, as it is a powerful tool to model and reason with uncertain and imprecision knowledge from low-quality PET images. In Chapter 2, we will introduce the fundamental background of the theory of belief functions (BFT). Then, BFT-based cancer treatment outcome prediction and automatic tumor segmentation will be discussed in following chapters.

Theory of Belief Functions

The theory of belief functions (BFT), also known as Dempster-Shafer or Evidence theory, was introduced by Dempster and Shafer [77, 82] and further elaborated by Smets [81, 83]. It is a generalization of both probability theory and set-membership approaches, and also closely relates to other methodologies, such as imprecise probability [84] or random sets [85]. As an effective theoretical framework for modeling, fusing, and reasoning with uncertain and/or imprecise information, the theory of belief functions has shown remarkable applications in divers fields [86], including data classification [87–97], data clustering [78, 79, 98–102], model parameter estimation [103–105], computer vision and image analysis [52, 106–112], and information fusion [80, 107, 113–117] etc..

In this chapter, the fundamental background will be described. As two main components of belief function theory, quantification of a piece of evidence and combination of different items of evidence will be recalled in Section 2.1 and Section 2.2, respectively. The issue of decision-making in the framework of belief functions will be discussed in Section 2.3. Then, some applications of belief functions that relate directly to our study in this thesis will be introduced in Section 2.4. We will conclude this chapter in Section 2.5.

2.1 Evidence Quantification

The theory of belief functions is a formal framework for reasoning under uncertainty based on the modeling of evidence [77]. Let ω be a variable taking values in a finite domain $\Omega = \{\omega_1, \cdots, \omega_c\}$, called the *frame of discernment*. An item of evidence regarding the actual value of ω can be represented by a *mass function* m on Ω, defined from the powerset 2^Ω to the interval $[0, 1]$, such that

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

Each number $m(A)$ denotes a *degree of belief* attached to the hypothesis that "$\omega \in A$". Function m is said to be normalized if $m(\emptyset) = 0$. Any subset A with $m(A) > 0$ is called a *focal element* of mass function m. If all focal elements are singletons, m is said to be *Bayesian*; it is then equivalent to a probability distribution. A mass function m with only one focal element is said to be *categorical* and is equivalent to a set.

Corresponding to a normalized mass function m, we can associate *belief* and *plausibility* functions from 2^{Ω} to $[0, 1]$ defined as:

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

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2.3}$$

Quantity $Bel(A)$ (also known as *credibility*) can be interpreted as the degree to which the evidence supports A, while $Pl(A)$ can be interpreted as the degree to which the evidence is not contradictory to A. Functions Bel and Pl are linked by the relation $Pl(A) = 1 - Bel(\overline{A})$. They are in one-to-one correspondence with mass function m, and they can be regarded as providing lower and upper bounds for the degree of belief that can be attached to each subset of Ω.

2.2 Evidence Combination

In the framework of belief functions, beliefs are elaborated by aggregating different items of evidence. The basic mechanism for evidence combination is *Dempster's rule* of combination [77]. *Dempster's rule of combination* [77], as well as its unnormalized version, i.e., the *conjunctive combination rule* defined in the transferable belief model (TBM) [81], are basic mechanisms for evidence fusion. Let m_1 and m_2 be two mass functions derived from independent items of evidence. They can be fused via the TBM conjunctive rule to induce a new mass function $(m_1 \textcircled{\cap} m_2)$ defined as

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{2.4}$$

This new mass function reduces uncertainty and imprecision via transferring masses of belief to conjunctions of the focal elements. Quantity $(m_1 \textcircled{\cap} m_2)(\emptyset)$ measures the *degree of conflict* between evidence m_1 and m_2. If $(m_1 \textcircled{\cap} m_2)(\emptyset) < 1$, the new mass function $m_1 \oplus m_2$

obtained by Dempster's rule can be represented as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-Q} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{2.5}$$

where $Q = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ measures the *degree of conflict* between evidence m_1 and m_2.

According to the value of Q, it can be found that when the conflict between m_1 and m_2 is significant, the combination result obtained by Dempster's rule becomes unreliable. To cope with this problem, Yager's rule [118] was designed to combine evidences with high contradiction, such as,

$$m(A) = \begin{cases} \sum_{B \cap C = A} m_1(B) m_2(C), & \text{when } A \subset \Omega; \\ m_1(\Omega) m_2(\Omega) + \sum_{B \cap C = \emptyset} m_1(B) m_2(C), & \text{when } A = \Omega; \\ 0, & \text{when } A = \emptyset; \end{cases} \tag{2.6}$$

Yager's rule transfers contradicting BBA to the frame of discernment. As the result, the uncertainty property of data is preserved when there is high conflicting between evidences.

Apart from Yager's rule, various other alternatives to Dempster's rule have also been developed under different situations, e.g., the TBM disjunctive combination rule [83], Dubois-Prade's rule [119], the weighted average [120, 121], and the cautious and bold disjunctive rules [116] etc.. In addition, the discounting strategy has been used in some other methods to deal with the conflicts, and a new dissimilarity measure consisting of both the conflict and distance has been introduced in [121] to determine the discounting factor of each source of evidence to be combined. The conflicts have also been used for detecting the change occurrences in the fusion of multi-temple information [109], like in the change detection of remote sensing. Nevertheless, these alternative methods usually increase the complexity for applications, Dempster's rule still remains the most popular one for combining independent evidence.

2.3 Decision Making

As has been discussed in Section 2.1, $\forall A \subseteq \Omega$, quantity $Bel(A)$ and $Pl(A)$ can be interpreted as the lower and upper bound of the belief that attached to the hypothesis "actual value is in A". Thus, plausibility and credibility can be adopted to define, respectively, the lower and upper expected risk of making a specific decision [122]. The two corresponding

decision rules are the *maximum plausibility rule* and the *maximum belief rule*, where the former one is a more optimistic strategy than the latter one.

Another commonly used decision rule transforms a mass function m into a probability function for decision-making. In the framework of transferable belief model (TBM) [81], the *pignistic probability transformation* is designed to transform m into the following probability distribution:

$$BetP(\omega_q) = \sum_{\omega_q \in A} \frac{m(A)}{|A|}, \tag{2.7}$$

for all $\omega_q \in \Omega$. It is actually a compromise between the maximum plausibility rule and the maximum belief rule.

2.4 Belief Functions in Data Classification And Clustering

Growing applications of the belief function theory have been reported in unsupervised learning [78,98,102,106], supervised learning [87,89,92,94,95], ensemble learning [123–125], and partially supervised learning [104,126], etc. It shows that learning using belief functions is getting more and more attention in statistical pattern recognition.

Among all the methods mentioned above, the Evidential K-NN (EK-NN) classification rule and the Evidential C-Means (ECM) clustering algorithm are two of the most representative learning methods based on belief functions. We will briefly introduce them in the following part.

2.4.1 Evidential K-NN Classification Rule

As the most representative classifier based on the theory of belief functions, an Evidential K-NN (EK-NN) classification rule was proposed in [87]. Depending on the informativeness of the training samples with respect to the class membership of the query pattern, the EK-NN classifier computes a mass function over the whole frame of classes, and provides a global treatment of imperfect training knowledge with uncertainty.

Let $\{(X_i, Y_i)|i = 1, \cdots, N\}$ be a collection of N training pairs, in which $X_i = [x_1, \cdots, x_V]^T$ is the ith training sample with V features and $Y_i \in \{\omega_1, \cdots, \omega_c\}$ is the corresponding class label. Given a query instance X^t, its class membership can be determined through the following steps:

- Each neighbor of X^t is considered as an item of evidence that supports certain

hypotheses regarding the class membership of X^t. Let X_j be one of its K nearest neighbors with class label $Y_j = \omega_q$. The mass function induced by X_j, which supports the assertion that X^t also belongs to ω_q is

$$\begin{cases} m_{t,j}(\{\omega_q\}) & = \alpha \exp(-\gamma_q d_{t,j}^2), \\ m_{t,j}(\Omega) & = 1 - \alpha \exp(-\gamma_q d_{t,j}^2), \end{cases} \tag{2.8}$$

where $d_{t,j}$ is the distance between X_j and X^t, while α and γ are two tuning parameters. According to the method presented [88], these parameters can be optimized via minimizing a performance criterion constructed on training data.

- Dempster's rule (2.5) is then executed to combine all neighbors' knowledge and obtain a global mass function for X^t. The lower and upper bounds for the belief of any specific hypothesis are then quantified via the credibility (2.2) and plausibility (2.3) values, respectively. In the case of {0,1} losses, the final decision on the class label of X^t can be made alternatively through maximizing the credibility, the plausibility, or the pignistic probability, as defined by Smets [81].

As an adaptive version of the EK-NN classifier, a neural network classifier based on the theory of belief functions has been proposed in [89]. Some other alternatives to the EK-NN method have also been developed. For instance, the credal classification methods [92–94] have been proposed by Liu et al. to deal with the overlapping classes in different cases. These methods permit the objects to be associated with not only the single classes but also meta-classes (i.e., disjunction of several classes) with different masses of belief, thus endowing the ability to specify the imprecision of classification.

2.4.2 Evidential C-Means

Let $\{X_1, \ldots, X_n\}$ be a collection of feature vectors in $\mathbb{R}^p$ describing n objects belonging to the set of clusters $\Omega = \{\omega_1, \cdots, \omega_c\}$. Evidential C-Means (ECM) is grounded on a new concept of partition, namely the *credal partition* [98], which extends the concepts of hard, fuzzy, and possibilistic partition by allocating, for each object, a mass of belief, not only to single clusters, but also to any subset of the whole frame Ω. Each single cluster ω_k, $k \in \{1, \ldots, c\}$, is represented by a prototype $V_k \in \mathbb{R}^p$. Then, for each nonempty subset $A_j \subseteq \Omega$, a centroid $\overline{V}_j$ is defined as the barycenter of the prototypes associated with the

singletons in A_j, i.e.,

$$\overline{V}_j = \frac{1}{c_j} \sum_{k=1}^{c} s_{kj} V_k, \tag{2.9}$$

where s_{kj} is binary, and it equals 1 iff $\omega_k \in A_j$; while $c_j = |A_j|$ denotes the cardinality of A_j.

Let $\mathbf{V}$ denotes a matrix of size $(c \times p)$ composed of the coordinates of the cluster centers such that V_{kq} is the qth component of the prototype V_k. ECM looks for a credal partition matrix $\mathbf{M} = (m_{ij})$ of size $(n \times 2^c)$ and for a matrix $\mathbf{V}$ by minimizing the following objective function:

$$\mathcal{J}_{ecm}(\mathbf{M}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} c_j^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}, \tag{2.10}$$

subject to the constraints $m_{ij} \geq 0$, $m_{i\emptyset} \geq 0$, and

$$\sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} m_{ij} + m_{i\emptyset} = 1, \quad \forall i = 1, \ldots, n, \tag{2.11}$$

where m_{ij} denotes the mass of the object X_i allocated to the credal cluster A_j; while $m_{i\emptyset}$ denotes that allocated to the empty set, and δ is a weighting parameter. The empty set is used for the detection of outliers. Coefficient $\alpha \geq 0$ controls the degree of penalization of the subsets according to their cardinality, and coefficient $\beta > 1$ controls the fuzziness of the credal partition.

Several variants of the ECM algorithms have also been proposed. For instance, a relational data version of ECM was proposed in [127]. A evidential clustering algorithm based on an alternative definition of the distance between a vector and the prototype of a meta-cluster was proposed in [100], which can produces more sensible results than the original ECM in situations where the prototype of a meta-cluster is close to that of singleton cluster. In [101], Zhou et al. introduce another variant of ECM, which is in fact an evidential counterpart to the median c-means and median fuzzy c-means algorithms.

2.5 Conclusion

In this chapter, we have briefly introduced the fundamental background of the theory of belief functions, including evidence quantification and fusion, decision-making rules, and the application of belief functions in data classification and clustering. In the following

chapters, our study focus on using the theory of belief functions to develop cancer treatment outcome prediction models and automatic tumor delineation algorithms.

Part II

Therapy Outcome Prediction based on Belief Functions and PET Images

Accurately predicting outcomes of radiation therapy is valuable for tailoring and adapting treatment planning. The goal of our study in this part is to develop reliable prediction models primarily using radiomic features extracted from FDG-PET images.

First, considering original feature space is usually high-dimensional containing unreliable input features, two alternative dimensionality reduction methods based on belief functions are proposed to improve the prediction performance:

- In Chapter 3, an evidential feature selection (EFS) method is proposed. Using a new fusion strategy to robustly quantify uncertainty and imprecision of studied datasets, EFS aims at sparsely selecting discriminant features for a modified evidential K-NN (EK-NN) classifier. The method proposed in this chapter has been published in: *C. Lian, S. Ruan, and T. Denœux,"An Evidential Classifier based on Feature Selection and Two-Step Classification Strategy",* Pattern Recognition*, Vol. 48, pages 2318-2327, 2015.*

- In Chapter 4, an evidential dissimilarity metric learning (EDML) method is proposed to realize a low-dimensional feature transformation and joint feature selection from input space, so as to improve both the accuracy and efficiency of the EK-NN classifier. The method presented in this chapter has been accepted for publication in: *C. Lian, S. Ruan, and T. Denœux, "Dissimilarity Metric Learning in the Belief Function Framework",* IEEE Transactions on Fuzzy Systems*, 2016 (in press, DOI: 10.1109/TFUZZ.2016.2540068).*

Second, to reliably predict treatment outcomes in radiation therapy, the imbalanced learning problem on small-sized datasets is carefully tackled in Chapter 5, since these two critical issues are frequently encountered in medical domain. The EFS method described in Chapter 3 is further improved by taking into account the imbalance and small-size of studied data. The work presented in Chapter 5 has been published in: *C. Lian, S. Ruan, T. Denœux, et al., "Selecting Radiomic Features from FDG-PET Images for Cancer Treatment Outcome Prediction",* Medical Image Analysis*, Vol. 32, pages 257-268, 2016.*

An Evidential Classifier Based on Feature Selection and Two-Step Classification Strategy

In this chapter, we investigate ways to learn efficiently from uncertain data using belief functions. In order to extract more knowledge from imperfect and insufficient information and to improve classification accuracy, we propose a supervised learning method composed of a feature selection procedure and a two-step classification strategy. Using training information, the proposed feature selection procedure automatically determines the most informative feature subset by minimizing an objective function. The proposed two-step classification strategy further improves the decision-making accuracy by using complementary information obtained during the classification process. The performance of the proposed method was evaluated on various synthetic and real datasets. A comparison with other classification methods is also presented.

3.1 Introduction

According to whether prior probabilities and class conditional densities are needed, supervised learning methods can be divided into two main categories, namely, parametric (model-based) and nonparametric methods [128]. Because they do not need any prior knowledge other than training samples, case-based classifiers (e.g., K-nearest neighbor rule [129], multilayer perceptrons [130], support vector machines [51] and decision trees [131]) are widely used in practice, and have proved to be very efficient. However, in the case of uncertain and imprecise data, many samples may be corrupted with noise or located in highly overlapping areas; consequently, it becomes difficult for these traditional methods

to obtain satisfactory classification results.

In this chapter, we explore two complementary ways to extract more useful knowledge from the training data:

- It often happens that the dataset contains irrelevant or redundant features. So as to efficiently learn from such imperfect training information, it is essential to find the most informative feature subset;
- Additional knowledge can be gained from the testing dataset itself to help reduce the possibility of misclassification. The "easy to classify" objects in the testing dataset can provide complementary evidence to help determine the specific class of the "hard to classify" objects.

To this end, a novel supervised learning method based on belief functions is proposed in this chapter. The proposed method is composed of a feature selection procedure and a two-step classification strategy, both based on a specific mass function construction method inspired by [132]. This method, called the "Dempster+Yager" combination rule, uses features of Dempster's rule, Yager's rule [118] and Shafer's discounting procedure [77] to achieve a better representation of uncertainty and imprecision in the EK-NN classifier. Through minimizing a new criterion based on belief functions, the proposed feature selection procedure searches for informative feature subsets that yield high classification accuracy and small overlap between classes. After feature selection, the proposed two-step classification strategy uses test samples that are easy to classify, as additional evidence to help classifying test samples lying in highly overlapping areas of the feature space.

The rest of this chapter is organized as follows. The proposed feature selection procedure and two-step classification strategy are discussed in Section 3.2. In Section 3.3, the proposed method is tested on different synthetic and real datasets, and a comparison with other methods is presented. Finally, conclusions are given in Section 3.4.

3.2 Proposed Method

Both the feature selection procedure and the two-step classification strategy proposed in this paper need proper handling of the uncertainty and imprecision in the data. To this end, a simple and specific mass function construction procedure will first be introduced in Section 3.2.1. The proposed feature selection procedure and two-step classification strategy

will then be presented, respectively, in Sections 3.2.2 and 3.2.3.

3.2.1 Construction of Mass Functions

We developed a specific combination rule to compute a mass function about the class label of a test sample, based on the evidence of its K-nearest neighbors. The proposed hybrid combination rule shares some features with Dempster's rule, Yager's rule [118] and Shafer's discounting procedure [77]. It will be referred to as the *"Dempster+Yager" rule* for short. In this rule, only singletons and the whole frame of discernment are considered as focal elements. Hence, all the imprecision will be succinctly represented by masses assigned to the whole frame of discernment.

As before, let $\{(X_i, Y_i), i = 1, \cdots, N\}$ be the training data. For an input instance X_t under test, the frame of discernment is $\Omega = \{\omega_1, \cdots, \omega_c\}$. Using the Dempster+Yager rule, the determination of X_t's mass function can be described as follows.

1. As in the classical E-KNN method [87], the K-nearest neighbors of X_t in the training set according to the Euclidean distance measure are first found. Let X_j be the jth nearest neighbor of X_t with $Y_j = \omega_q$. The evidence regarding X_t's class label provided by X_j is quantified as described by (2.8).

2. Nearest neighbors with the same class label ω_q are then grouped in a set Γ_q ($q = 1, \ldots, c$). As the mass functions in the same set Γ_q have the same focal elements, there is no conflict between them. So, regardless of outliers (a particular situation that is not considered in our approach), Dempster's rule is appropriate to combine the pieces of evidences in Γ_q. As a result, the evidence provided by nonempty Γ_q is represented as a simple mass function,

$$m_t^{\Gamma_q}(\{\omega_q\}) = 1 - \prod_{j \in \Gamma_q} m_{t,j}(\Omega), \tag{3.1a}$$

$$m_t^{\Gamma_q}(\Omega) = \prod_{j \in \Gamma_q} m_{t,j}(\Omega). \tag{3.1b}$$

   If Γ_q is empty, then $m_t^{\Gamma_q}$ is defined as the vacuous mass function defined by $m_t^{\Gamma_q}(\Omega) = 1$;

3. When most neighbors of a testing instance X_t belong to a specific class (e.g., ω_q), the degree belief that X_t also belongs to this class should be large. Consequently, we can postulate that the reliability of the evidence provided by each set Γ_q is increasing

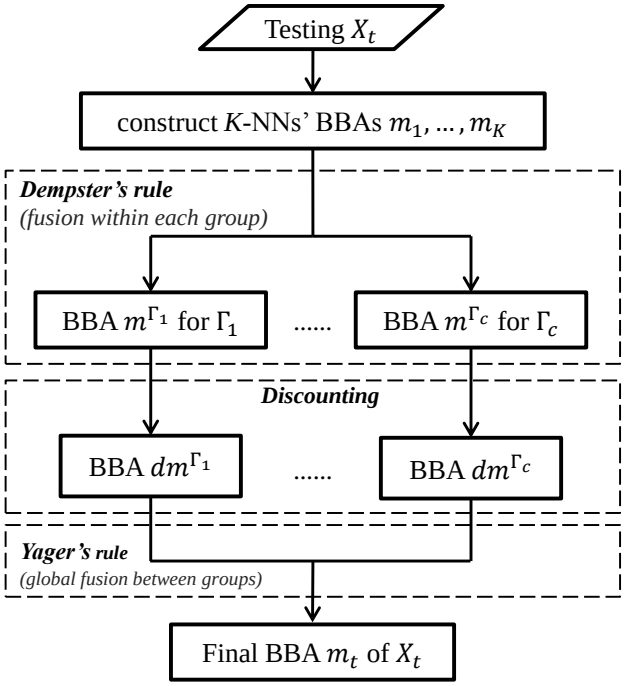


Figure 3.1: Flowchart of mass function construction. Mass functions m^{Γ_q}, dm^{Γ_q} for $q = 1, \ldots, c$ and m_t are calculated by (3.1) to (3.3).

with its cardinality $|\Gamma_q|$. The mass functions obtained in last step should thus be further discounted as

$$dm_t^{\Gamma_q}(\{\omega_q\}) = \left(\frac{|\Gamma_q|}{|\Gamma_{max}|}\right)^{\eta} m_t^{\Gamma_q}(\omega_q), \tag{3.2a}$$

$$dm_t^{\Gamma_q}(\Omega) = 1 - \left(\frac{|\Gamma_q|}{|\Gamma_{max}|}\right)^{\eta} m_t^{\Gamma_q}(\omega_q), \tag{3.2b}$$

where $|\Gamma_{max}|$ is the maximum cardinality within $\{|\Gamma_1|, \cdots, |\Gamma_c|\}$, and $\eta \geq 0$ is a coefficient that controls the discounting level. A larger value of η results in stronger discounting. In particular, when $\eta = 0$, there is no discounting at all. The value of η can be determined by minimizing the leave-one-out cross-validation error rate. Generally, good results are obtained if we take $\eta \in [0, 2]$.

4. After the discounting procedure described in the previous step, the mass functions at hand may still be partially conflicting, especially when there are similar numbers of nearest neighbors with different class labels. Since Yager's rule can have a better behavior than Dempster's rule when combining highly conflicting evidences [118,133], it is chosen at this step to fuse the probably conflicting mass functions in sets Γ_1 to Γ_c obtained in the previous step. As the result, the global mass function regarding

the class label of object X_t is finally given by

$$m_t(\{\omega_q\}) = dm_t^{\Gamma_q}(\omega_q) \prod_{h\in\{1,\dots c\}\setminus q} dm_t^{\Gamma_h}(\Omega), \quad q = 1,\dots,c, \tag{3.3a}$$

$$m_t(\Omega) = 1 - \sum_{q=1}^{c} \left(dm_t^{\Gamma_q}(\{\omega_q\}) \prod_{h\in\{1,\dots c\}\setminus q} dm_t^{\Gamma_h}(\Omega) \right), \tag{3.3b}$$

The focal elements of m_t are singletons and the whole frame of discernment. Consequently, the credibility and plausibility criteria (i.e., Bel_t and Pl_t) will lead to the same hypotheses about X_t.

The mass function construction procedure discussed above is summarized as a flowchart in Figure 3.1. It combines the advantages of Dempster's and Yager's rules. Hence, in classification applications, this specific procedure allows for a more robust representation of uncertainty than that obtained using any of the two classical combination rules. To better illustrate the performance of the proposed Dempster+Yager rule, two examples are given below.

Table 3.1: Combination result with different rules in Example 1.

	Neighbors			Dempster's rule	Yager's rule	Dempster+Yager rule
	♯ 1	♯ 2	♯ 3			
$m(\{\omega_1\})$	0.8	0.8	0	0.8276	0.1920	**0.7680**
$m(\{\omega_2\})$	0	0	0.8	0.1379	0.0320	**0.0080**
$m(\Omega)$	0.2	0.2	0.2	0.0345	0.7760	**0.2240**

Example 1. *To simulate a situation with conflicting pieces of evidence, we let the number of nearest neighbors be $K = 3$, and we assume that the test sample X_t lies at the same distance to all the three nearest neighbors. The first two neighbors of X_t belong to class ω_1, and the third one belongs to class ω_2. We assume that $\Omega = \{\omega_1, \omega_2\}$ and $\eta = 2$. The three mass functions and the result of their combination by Dempster's rule, Yager's rule and our Dempster+Yager rule are shown in Table 3.1. In this case, the Dempster+Yager rule is more conservative than Dempster's rule (it assigns a larger mass to Ω), while being more specific than Yager's rule.*

Example 2. *Table 3.2 illustrates an even more conflicting situation, in which two neighbors belong to ω_1 and two neighbors belong to ω_2. We still assume that the test sample X_t is at the*

Table 3.2: Combination result with different rules in Example 2.

	Neighbors				Dempster's rule	Yager's rule	Dempster+Yager rule
	♯1	♯2	♯3	♯4			
$m(\{\omega_1\})$	0.8	0.8	0	0	0.4898	**0.0384**	**0.0384**
$m(\{\omega_2\})$	0	0	0.8	0.8	0.4898	**0.0384**	**0.0384**
$m(\Omega)$	0.2	0.2	0.2	0.2	0.0204	**0.9232**	**0.9232**

same distance to all nearest neighbors, and we take $\eta = 2$. *In this case, the Dempster+Yager rule yields the same result as Yager's rule. Both rules assign a large mass to the whole frame of discernment.*

To sum up, the proposed Dempster+Yager rule can lead to more reliable quantification of uncertainty than the original Dempster's rule and Yager's rule at the above two different situations.

3.2.2 Evidential Feature Selection

In pattern recognition applications, the data may contain irrelevant or redundant features. Feature selection techniques are intended to cope with this issue. They aim to select a subset of features that can facilitate data interpretation while reducing storage requirements and improving prediction performance [134]. Filter, wrapper and embedded methods are three main categories of algorithms that are widely used for feature selection [135]. Filter methods such as described in [136–138], which use variable ranking as the principal selection mechanism, are simple and scalable. However, they may produce a sub-optimal subset because they do not take into account the correlation between features [134]. In contrast, wrapper and embedded methods, such as sequential selection algorithms [139, 140] and direct objective optimization methods [141], use the prediction accuracy of given classifiers as the criterion for selecting feature subset. They are more likely to find optimal feature subsets than filter methods. However, up to now, none of the available wrapper or embedded methods were designed to work for imperfect data with high uncertainty and/or imprecision. Such a feature selection procedure, called evidential feature selection (EFS), is proposed in this section.

The proposed method tackles the feature selection issue from a novel perspective. It aims to meet the following three requirements:

1. The selected features should be informative regarding the class labels, i.e., they should not yield lower classification accuracy than the complete set of features;

2. The selected feature subset should have the ability to reduce the uncertainty of the data, i.e., it should result in a small overlap between different classes in the feature space;

3. The selected features should be as sparse as possible. A feature subset with smaller cardinality implies lower storage requirement and lower risk of overfitting.

The above three requirements can be met simultaneously by minimizing an objective function derived from the training samples. In order to present this objective function clearly, a simple form of weighted Euclidean distance should be discussed at first. Depending on the values of a binary coefficient vector, this weighted Euclidean distance will generate different sets of K nearest neighbors for a sample under test. The weighted distance between a test sample X^t and a training sample X_i with m features is proposed as

$$d_{t,i} = \sqrt{\sum_{p=1}^{m} \lambda_p (d_{t,i}^p)^2}, \tag{3.4}$$

where $d_{t,i}^p$ $(1 \leq p \leq m)$ is the difference between the values of the pth components of the two feature vectors and $\lambda_p \in \{0, 1\}$ is the corresponding coefficient to be determined. Obviously, the feature subset can be selected by changing the values of the coefficient vector. As the result, the pth component of the feature vector will be selected when $\lambda_p = 1$ and it will be eliminated when $\lambda_p = 0$.

Based on the weighted Euclidean distance measure (3.4), and using the mass function construction procedure introduced in Section 3.2.1, we can propose an objective function satisfying the above three requirements for a qualified feature subset. Let $\{(X_i, Y_i), i = 1, \cdots, N\}$ be a training set. The proposed three-term objective function is defined as

$$obj = \frac{1}{n} \sum_{i=1}^{n} \sum_{q=1}^{c} (Pl_i(\omega_q) - t_{i,q})^2 + \frac{\rho}{n} \sum_{i=1}^{n} m_i(\Omega) + \delta \sum_{p=1}^{m} [1 - \exp(-\mu \lambda_p)]. \tag{3.5}$$

In (3.5), the first term is a squared error corresponding to the first requirement discussed above, Pl_i is the plausibility function of training sample X_i and $t_{i,q}$ is the qth component of a c-dimensional binary vector t_i such that $t_{i,q} = 1$ if $Y_i = \omega_q$ and $t_{i,q} = 0$ otherwise. The second term is the average mass assigned to the whole frame of discernment. It penalizes feature subsets that result in high uncertainty and imprecision, thus allowing us to meet

the second requirement. The last term, which is an approximation of the l_0-norm as used in [142], forces the selected feature subset to be sparse. Here, ρ and δ are two hyper-parameters in $[0, 1]$, which influence, respectively, the number of uncertainty samples and the sparseness of resulting feature subset. Their values should be tuned to maximize the classification accuracy. Coefficient μ is kept constant; according to [142], it is often set to 5.

Using (2.8)-(3.3), the objective function (3.5) can be written as

$$obj = \frac{1}{n}\sum_{i=1}^{n}\sum_{q=1}^{c}\left(1 - t_{i,q} - \sum_{h\neq q} B_h^i\right)^2 + \frac{\rho}{n}\sum_{i=1}^{n}\left(1 - \sum_{q=1}^{c} B_q^i\right) + \delta\sum_{p=1}^{m}[1 - \exp(-\mu\lambda_p)], \quad (3.6)$$

with

$$B_q^i = A_q^i \prod_{s\in\{1,\dots c\}\setminus q}(1 - A_s^i) \quad (3.7)$$

and

$$A_q^i = \left(\frac{|\Gamma_q^i|}{|\Gamma_{max}^i|}\right)^{\eta}\left(1 - \prod_{j\in\Gamma_q^i}\left[1 - \alpha\exp(-\gamma_q \cdot d_{i,j}^2)\right]\right), \quad (3.8)$$

where $d_{i,j}$ is the distance between the training sample X_i and its jth nearest neighbor computed using (3.4) ,with coefficients $\{\lambda_1, \cdots, \lambda_c\}$ to be optimized. During the optimization process, the K nearest neighbors for each training sample (X_i, Y_i) are determined by the weighted distance measure (3.4) with the current weights $\{\lambda_1, \cdots, \lambda_c\}$. The mass functions m_i are computed using the construction procedure presented in Section 3.2.1, followed by the calculation of the plausibility value Pl_i using (2.3). Mass and plausibility values change with binary coefficients $\{\lambda_1, \cdots, \lambda_c\}$, which finally drives the decrease of the objective function (3.5)-(3.6).

As a global optimization method, the integer genetic algorithm [143, 144] can properly solve the integer optimization problem without gradient calculation. Hence, it is chosen in this paper to optimize $\{\lambda_1, \cdots, \lambda_c\}$, so as to find a good feature subset.

3.2.3 Two-Step Classification

After selecting features using the procedure described in the previous section, a two-step classification strategy allows us to classify unknown test samples based on belief functions.

For a test dataset $\mathbf{T} = \{S_j, j = 1, \ldots, n_t\}$, this two-step classification strategy can be described as follows:

Step 1 Using the Dempster+Yager combination rule, the mass function m_j of each test sample S_j is first derived from training pairs (X_i, Y_i), $i = 1, \ldots, N$. Based on m_j, the collection $\mathbf{T}$ is divided into two groups T^1 and T^2, where $\mathrm{T}^1 = \{S_j : \max_{A \subseteq \Omega} m_j(A) \neq m_j(\Omega)\}$ and $\mathrm{T}^2 = \{S_j : \max_{A \subseteq \Omega} m_j(A) = m_j(\Omega)\}$;

Then, test samples in T^1 are classified into the classes with highest masses. For instance, if $m_j(\{\omega_1\}) > m_j(\{\omega_q\})$ for all $q \neq 1$, we label S_j as ω_1;

Step 2 After classifying the test samples in T^1, we add these labeled test samples to the training set $\{(X_i, Y_i), i = 1, \ldots, N\}$, and therefore obtain a larger training set $\{(X_i', Y_i'), i = 1, \ldots, N'\}$. The center (or prototype) p_j of each class ω_j is then defined by averaging the training samples corresponding to this class,

$$p_j = \frac{1}{c_j} \sum_{Y_i' = \omega_j} X_i', \tag{3.9}$$

where c_j is the cardinality of the set $\{X_i' | Y_i' = \omega_j\}$ of training patterns in class ω_j, and $j = 1, \ldots, c$.

To each test pattern in group T^2 (i.e., uncertain samples with the largest mass of belief on Ω), the Mahalanobis distance measure is adopted to compute the distances of this test pattern to each class center. As compared to the standard Euclidean distance, this metric can more effectively take into account the correlations between different samples in studied datasets. Let S_0 be a test sample within T^2, the distance from it to center p_j is

$$md(S_0, p_j) = \sqrt{\sum_{q=1}^{m} \frac{(S_0^q - p_j^q)^2}{(\delta_j^q)^2}}, \tag{3.10}$$

where S_0^q and p_j^q are, respectively, the qth dimension of S_0 and p_j, and δ_j^q is the standard deviation of the qth feature among training samples belonging to class ω_j. Based on the distances $\{md(S_0, p_1), \cdots, md(S_0, p_m)\}$, S_0 is finally allocated to the nearest class.

Using the procedure discussed above, test samples that are easy to classify provide additional evidence to help classifying highly uncertainty test samples. As will be shown in the next section, this strategy enhances the classification accuracy of the EK-NN rule, especially in highly overlapping regions of the feature space.

3.3 Experimental Results

The presented experiments are composed of two parts. In the first part, the feasibility of the proposed feature selection procedure was evaluated on two synthetic datasets. In each synthetic dataset, the numbers of relevant, redundant and irrelevant features were varied to assess the robustness of the method under different situations. In addition, to show the validity of the two-step classification strategy, we compared it in detail with the EK-NN classifier [87, 88, 128] on another synthetic dataset.

In the second part, we first compared the performance of the proposed feature selection procedure with some classical wrapper selection methods on seven real datasets. Then, on the same real datasets, the classification accuracy of the proposed two-step classification strategy was compared with other well-known classifiers after selecting features using different methods. Finally, we tried to determine whether the proposed feature selection procedure can help to improve classification performance of other classifiers. The classification performance of the proposed two-step procedure was further compared with other methods using the same feature subsets selected by the proposed procedure.

3.3.1 Performance on Synthetic Datasets

3.3.1.1 Feature Selection

The feasibility of the proposed feature selection procedure was assessed on two different kinds of synthetic datasets. The generating mechanisms for the two different datasets are described below.

Synthetic Data 1 These data were generated using the procedure presented in [145]: The feature space contains n_r informative features uniformly distributed between -1 and +1. The output label for a given sample is defined as

$$y = \begin{cases} \omega_1 & \text{if } \max_i(x_i) > 2^{1-\frac{1}{n_r}} - 1, \\ \omega_2 & \text{otherwise,} \end{cases} \tag{3.11}$$

were x_i is the ith feature. Besides the relevant features, we added n_i irrelevant features uniformly distributed between -1 and +1, without any relation with the class label; and n_c redundant features copied from the relevant features. The optimal discriminating surface for this synthetic data is highly non-linear.

Table 3.3: Cardinality of selected feature subsets for synthetic data 1, and comparison of classification error (in %) between selected feature subset (with EFS) and all features (without EFS). Here n_r, n_c and n_i represent the number of relevant, redundant and irrelevant features, respectively.

n_r	n_c	n_i	Subset size	EK-NN error	Two-step classification error	
					Without EFS	With EFS
2	2	6	2	14.67	12.67	**2.67**
2	2	16	2	17.33	12.00	**1.33**
2	2	26	2	23.33	18.67	**4.00**
2	2	36	2	28.67	26.67	**5.33**
2	2	46	2	29.33	23.33	**4.67**

Synthetic Data 2 To generate data, two informative features were first obtained from four different two-dimensional normal distributions, $N(m_1, I)$ and $N(m_2, I)$ for class 1; $N(m_3, I)$ and $N(m_4, I)$ for class 2. Here, $m_1 = [3, 3]$, $m_2 = [6, 6]$, $m_3 = [3, 6]$ and $m_4 = [6, 3]$. In addition, we added n_i irrelevant features, all randomly generated from the normal distribution $N(4.5, 2)$, and n_c redundant features copied from relevant features.

For both synthetic datasets, we set $n_r = 2$, $n_i \in \{6, 16, 26, 36, 46\}$ and $n_c = 2$ to simulate five different situations. In each case, we generated 150 training instances, and used the proposed procedure to search for the most informative feature subset. Then, 150 test instances were generated. We used the EK-NN classifier to classify these test instances with all features, and simultaneously used the proposed two-step classification strategy to classify them with all features and with the selected feature subset. In the five situations, we always set $\eta = 0.5$, $\rho = 0.5$, $\delta = 0.05$ and $K = 5$. The results are shown in Tables 3.3 and 3.4. For both datasets, the selection procedure always found the two relevant features. The two-step classification strategy resulted in higher accuracy than the EK-NN classifier. The feature selection procedure brought further improvement of classification performance, especially when the dimension of the initial feature space was high. These results show the good performance provided by the proposed feature selection procedure.

Table 3.4: Cardinality of selected feature subsets for synthetic data 2, and comparison of classification error (in %) between selected feature subset (with EFS) and all features (without EFS). Here n_r, n_c and n_i represent the number of relevant, redundant and irrelevant features, respectively. The number of relevant features here is two (i.e., $n_r = 2$).

n_c	n_i	Subset size	EK-NN error	Two-step classification error	
				Without EFS	With EFS
2	6	2	21.33	12.00	**8.67**
2	16	2	34.67	26.00	**14.67**
2	26	2	31.33	27.33	**16.00**
2	36	2	52.67	37.33	**11.33**
2	46	2	50.00	39.33	**8.00**

3.3.1.2 Two-Step Classification

In addition to the previous experiment, the performance of the proposed two-step classification strategy was tested solely on another synthetic dataset constructed from four normal distributions with means $m_1 = [3, 3]$, $m_2 = [3, 6.5]$, $m_3 = [6.5, 3]$, $m_4 = [6.5, 6.5]$ and variance matrix $\Sigma = 2I$. Instances generated from $N(m_1, \Sigma)$ and $N(m_2, \Sigma)$ with equal probabilities were labeled as ω_1, while other instances generated from $N(m_3, \Sigma)$ and $N(m_4, \Sigma)$ with equal probabilities were labeled as ω_2. Classes ω_1 and ω_2 had the same number of instances, and the sizes of training and testing datasets were both 500.

The classification results of the two-step classification strategy were compared with those of the EK-NN classifier with $K = 5$ and $\eta = 0.5$. Figure 3.2(a) shows the training samples and the corresponding test samples. Figures 3.2(b) and (c) display the credal partitions (i.e., the mass functions for each of the test samples [78, 98]) obtained, respectively, using the EK-NN classifier and the proposed method. The blue, green and black points represent instances with highest mass function on $\{\omega_1\}$, $\{\omega_2\}$ and Ω, respectively. When comparing Figures 3.2(b)-(c) with Figure 3.2(a), we can see that the proposed method results in more imprecise mass functions for the test samples in overlapping regions. This is mainly because the proposed Dempster+Yager rule has better ability than Dempster's rule to deal with highly imprecise instances (such as the boundary samples shown in Figure 3.2(c)).

Figures 3.2(d)-(f) show the classification results obtained, respectively, by EK-NN, the Dempster+Yager rule and the two-step classification strategy; the magenta stars represent

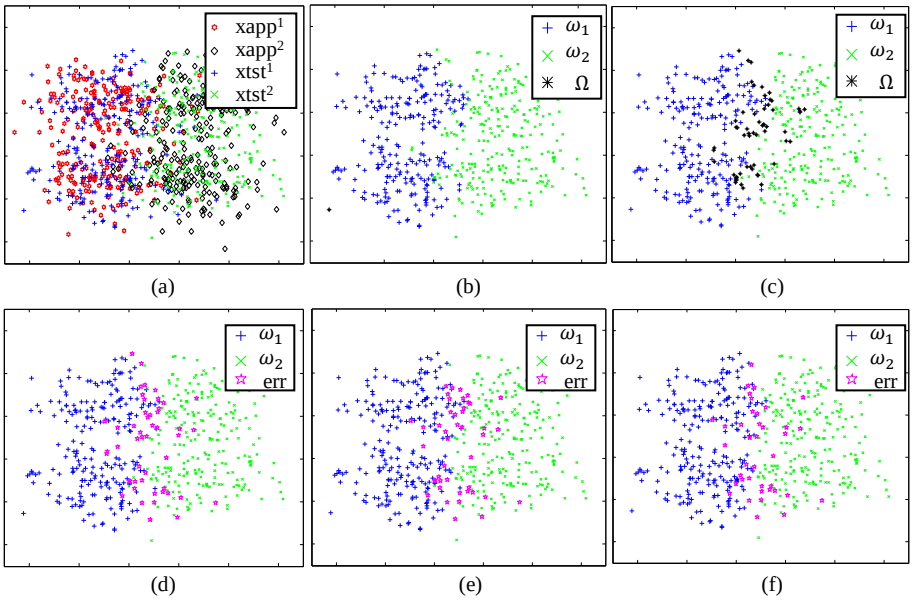


Figure 3.2: Test of the two-step classification strategy on a synthetic dataset; (a) shows training and test samples; (b) and (c) are credal partition obtained, respectively, by the EK-NN classifier and the two-step classification rule. The blue, green and black points represent instances with highest mass function on $\{\omega_1\}$, $\{\omega_2\}$ and Ω respectively; (d)-(f) are classification results obtained, respectively, by EK-NN, the proposed Dempster+Yager combination and the two-step classification strategy; the magenta stars represent misclassification instances. The calculated error rates for (d)-(f) are, respectively, 9.80%, 8.80% and 7.80% (color version is suggested).

misclassified instances. These results show that the proposed Dempster+Yager combination rule yields higher classification accuracy than EK-NN on these imprecise data and the two-step classification strategy further improves the performance. The calculated error rates for EK-NN, Dempster+Yager combination rule and two-step classification strategy are, respectively, 9.80%, 8.80% and 7.80%.

Table 3.5: Influence of parameter η on the proposed method.

η	0	0.5	1	1.5	2
Error rate (%)	11.03	**10.94**	11.26	11.27	11.27

In addition, we also estimated the influence of parameter η on our two-step classification procedure, using this synthetic dataset. The value of η was chosen in$\{0, 0.5, 1, 1.5, 2\}$, K was set to 5, and we evaluated the performance 50 times with each η. The average misclassification error rates are reported in Table 3.5. As can be seen, the value of η had

Table 3.6: Briefly description of the seven real datasets used in our experiments.

Dataset	Class amount	Feature amount	Sample amount
Iris	3	4	150
Seeds	3	7	210
Wine	3	13	178
Yeast	3	8	1055
WDBC	2	30	569
Parkinsons	2	22	195

some limited influence on the classification accuracy, although the procedure appears not to be very sensitive to this coefficient. The best performance was obtained with $\eta = 0.5$.

3.3.2 Performance on Real Datasets

In this section, the proposed feature selection procedure and two-step classification strategy are compared with some classical wrapper selection methods and usual classifiers. The comparison was performed on six real datasets downloaded from the UCI Machine Learning Repository [146]. Some characteristics of these datasets are summarized in Table 3.6. As in [93], "in the yeast dataset, three classes named as CYT, NUC and ME3 were selected, since these three classes are close and difficult to discriminate".

3.3.2.1 Feature Selection Performance

The proposed feature selection procedure was compared with three classical wrapper methods: sequential forward selection (SFS), sequential backward selection (SBS) and sequential floating forward selection (SFFS) [135, 139]. We used ten-fold cross validation for the six UCI datasets. For all datasets, we iteratively chose one subset of the data as the test set, and treated the other subsets of data as training samples. At each iteration, we used SFS, SBS, SFFS and the proposed EFS to select features from the training data, and then executed the proposed two-step classification strategy to classify test instances with the selected feature subsets. The average misclassification rates obtained by different methods were calculated. In addition, based on feature frequency statistics, the robustness of selected feature subsets was evaluated using the method introduced in [1].

Table 3.7: Comparison of the proposed feature selection method with classical wrapper methods on seven real datasets. The proposed two-step classification was used to obtain average misclassify ratio. The robustness of selected feature subset is evaluated by the way proposed in [1].

	Iris			**Seeds**		
	Error (%)	Robustness (%)	Subset size	Error (%)	Robustness (%)	Subset size
All	2.67	n/a	4	7.62	n/a	7
SFS	4.67	54.55	1	11.90	57.97	2
SBS	5.33	21.05	2	10.95	23.88	3
SFFS	5.33	21.62	3	5.24	54.93	2
EFS*	***2.00***	***100***	3	***4.76***	***81.18***	3
	Wine			**Yeast**		
	Error (%)	Robustness (%)	Subset size	Error (%)	Robustness (%)	Subset size
All	13.04	n/a	13	38.87	n/a	8
SFS	30.50	75	1	61.99	***100***	1
SBS	6.24	42.47	5	48.35	***100***	1
SFFS	7.29	57.58	4	36.21	40	5
EFS*	***5.13***	***91.89***	3	***32.51***	***100***	2
	WDBC			**Parkinsons**		
	Error (%)	Robustness (%)	Subset size	Error (%)	Robustness (%)	Subset size
All	7.20	n/a	30	13.37	n/a	22
SFS	14.44	80	1	15.82	33.33	1
SBS	19.67	22.22	2	19.03	23.91	2
SFFS	9.87	25	4	13.79	43.65	3
EFS*	***5.80***	***92.37***	3	***8.63***	***100***	3

The misclassification rate, robustness and average feature subset size for all methods are summarized in Table 3.7. As can be seen, the proposed feature selection procedure performed uniformly well on all datasets. It resulted in more robust feature subsets than the other three classical wrapper methods, and simultaneously yielded higher classification accuracy.

Table 3.8: Misclassification rates (in %) of the proposed method and six other classifiers obtained by 10-fold cross-validation. For BK-NN and CCR, R_e and R_i represent, respectively, the error and imprecision rates. Both the proposed EFS and the classical SFFS have been used to select feature for the six compared classifiers.

		Iris	**Seeds**	**Wine**	**Yeast**	**WDBC**	**Parkinsons**
SFFS +	**ANN**	8.00	7.62	9.64	32.57	9.15	9.63
	CART	8.00	7.14	9.09	37.55	10.04	11.21
	SVM	6.00	7.14	6.83	36.14	8.28	13.26
	EK-NN	5.33	6.67	6.18	35.07	9.70	16.39
	BK-NN (R_e, R_i)	(4.00,4.67)	(2.38,11.90)	(6.74,5.13)	(16.31,40.84)	(7.22,8.44)	(9.18,11.37)
	CCR (R_e, R_i)	(4.00,4.67)	(3.81,18.57)	(3.99,15.33)	(19.53,36.11)	(5.99,15.83)	(16.42,12.26)
EFS +	**ANN**	5.33	***4.76***	6.18	33.84	6.32	12.35
	CART	7.33	7.62	6.71	36.78	7.56	11.82
	SVM	4.67	5.24	5.60	32.71	6.33	11.38
	EK-NN	4.00	5.71	***4.45***	37.05	5.98	10.69
	BK-NN (R_e, R_i)	(2.00,4.67)	(3.33,10.00)	(2.19,6.22)	(16.95,40.77)	(3.69,7.73)	(5.58,15.35)
	CCR (R_e, R_i)	(2.67,3.33)	(10.48,6.19)	(3.93,5.07)	(31.66,8.82)	(4.19,15.01)	(17.49,8.58)
EFS + Two-step		***2.00***	***4.76***	5.13	***32.51***	***5.80***	***8.63***

3.3.2.2 Classification performance

Still using the six real datasets presented in Table 3.6, the classification performance of the proposed two-step classification was compared with that of six other classifiers: Artificial Neural Networks (ANN) [147], Classification And Regression Tree (CART) [131], Support Vector Machine (SVM) [51], EK-NN, Belief-based K-Nearest neighbor classifier

(BK-NN) [92] and CCR [93]. The first three methods are classical classifiers, while the last three are either well-known or recent evidential classifiers based on belief functions. We can remark that, in BK-NN and CCR, the classification performance is assessed using two measures: the error rate $R_e = (N_e/T) \times 100\%$, where N_e is the number of misclassified samples assigned to wrong meta-classes, and T is the number of test samples; and the imprecision rate $R_I = (N_I/T) \times 100\%$, where N_I is the number of test samples with highest mass functions on non-singletons (i.e., on meta-classes). The BK-NN and CCR methods do not make any direct decision for highly imprecise samples, but transfer them to the meta-classes. Hence, the error rate R_e of BK-NN and CCR is decreased.

Both a classical wrapper selection method, i.e., sequential floating forward selection (SFFS), and the proposed EFS were used with other six classifiers. The classification performance of these classifiers were then compared with that of the proposed two-step classification integrating EFS. As in the previous experiment, the 10-fold cross-validation was adopted to quantify classification results. The average misclassification rates obtained by different classifiers are reported in Table 3.8. As can be seen, the proposed method performs better than ANN, CART, SVM and EK-NN in this experiment. BK-NN and CCR resulted in the lowest error rate on the Seeds and Wine data. However, due to the fact that a nonspecific decision has been made for uncertain objects, they also have large imprecision rates. Therefore, we can conclude that the proposed classification method performed well on these real datasets.

3.3.2.3 Generality of The Proposed Method

From Table 3.8, we can also find that, as compared to SFFS, the proposed EFS further improved the performance of testing classifiers in most cases. These results show that EFS is in some sense general as it can be used with other classifiers. However, it works better if it is used for the proposed two-step classification.

To further evaluate the generality of the proposed EFS, using the same feature subsets selected by it, we compared the classification performance of the proposed two-step classification with that of other classifiers. In order to make the comparison more comprehensive, we used 2-fold cross-validation in this test, so as to simulate a situation in which there are more test data but less training data. The comparison was executed 200 times. The average error rates for the different classifiers are reported in Table 3.9. As can be seen, all classifiers performed poorly on the Yeast data. This dataset is actually very difficult to

Table 3.9: Misclassification rates obtained by 2-fold cross-validation for different classifiers using the same feature subsets selected by the proposed EFS.

	Iris	**Seeds**	**Wine**	**Yeast**	**WDBC**	**Parkinsons**
ANN	6.23	8.62	7.07	35.09	6.52	13.92
CART	5.50	11.70	8.22	37.76	8.07	16.75
SVM	3.78	9.72	5.71	33.68	6.47	13.53
EK-NN	4.04	6.19	5.96	38.20	***5.71***	12.43
BK-NN (R_e, R_i)	(2.03,5.67)	(3.96,7.44)	(4.57,6.67)	(18.92,40.03)	(5.97,7.19)	(9.29,16.03)
CCR (R_e, R_i)	(3.49,2.90)	(5.79,16.73)	(5.01,3.72)	(20.88,38.52)	(6.83,5.39)	(19.28,5.55)
Two-step	***2.52***	***4.94***	***4.42***	***32.97***	5.86	***12.37***

classify. The BK-NN and CCR methods yielded lower error rates than did our method on these data. However, due to the fact that nonspecific decisions can be made for uncertain objects, they also yielded large imprecision rates. Similar results can be found on the Iris and Seeds data when comparing BK-NN with our method. On the WDBC and Parkinsons data, EK-NN and the proposed two-step classification had similar performance.

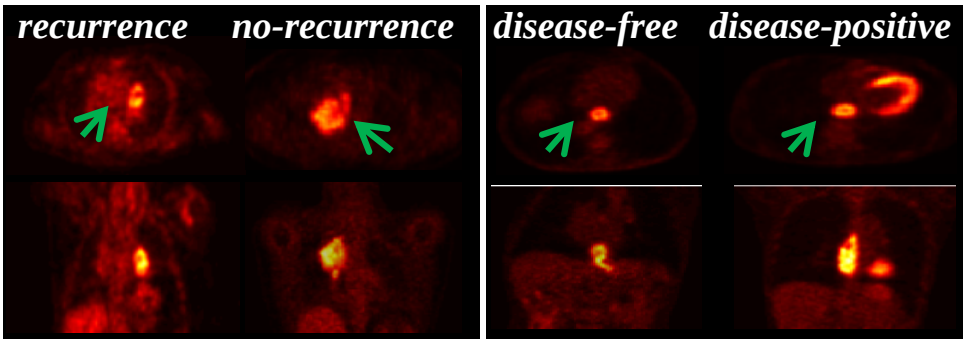
In summary, it appears from these results that the proposed two-step classification generally outperformed the other classifiers on the real datasets considered in these experiments. The proposed feature selection procedure has also been found to yield better results when used jointly with the proposed two-step classification strategy.

3.3.3 Performance on Clinical Datasets

In this experiment, the proposed method is evaluated by two real patient datasets:

1) Lung Tumour Data: Twenty-five patients with stage II-III non small cell lung cancer were studied. 52 SUV-based (SUV_{max}, SUV_{mean}, SUV_{peak}, MTV and TLG) and texture-based (gray level size zone matrices (GLSZM) [20]) features were extracted. The definition of recurrence for patients at one year after the treatment is primarily clinical with biopsy and PET/CT. Local or distant *recurrence* is diagnosed on 19 patients, while *no recurrence* is reported on the remaining 6 patients (example images can be seen in Figure 3.3(a)).

2) Esophageal Tumor Data: Thirty-six patients with esophageal squamous cell carci-



(a) Lung Tumors (b) Esophageal Tumors

Figure 3.3: Examples of tumor uptakes on FDG-PET imaging from different views; (a) recurrence and no-recurrence instances before treatment of lung tumor; (b) disease-free and disease-positive instances before treatment of esophageal tumor.

nomas were studied. We have 29 SUV-based (SUV_{max}, SUV_{mean}, SUV_{peak}, MTV and TLG), GLSZM-based and patients' clinical features (gender, tumour stage and location, WHO performance status, dysphagia grade and weight loss from baseline). The disease-free evaluations include a clinical examination with PET/CT and biopsies. 13 patients were labeled *disease-free* when neither loco regional nor distant tumor recurrence is detected, while the remaining 23 patients were diagnosed as *disease-positive* (example images can be seen in Figure 3.3(b)).

The detailed description of the above datasets and extracted features can be found in Chapter 5.

3.3.3.1 Feature Selection Performance

In the leave-one-out cross-validation (LOOCV) protocol, the proposed evidential feature selection (EFS) was compared with two classical wrapper methods (SFS and SFFS [139]) and a widely used imbedded method, SVMRFE [141]. The classification accuracy of SVM (Gaussian kernel with $\sigma = 1$ was empirically chosen) serves as the selection criteria in SFS and SFFS. For the proposed EFS, an integrated LOOCV on training set was adopted to tune the corresponding hyper-parameters. The cutoff thresholds for all the last three methods (feature subsets selection) were determined as that obtained best prediction performance. In each iteration of the exterior LOOCV, the selected feature subsets were used to predict the test data. The same SVM classifier was still used after SFS and SFFS, while the modified EK-NN discussed in Section 3.2.1 was executed after EFS. Finally, the

Table 3.10: Comparing feature selection methods using leave-one-out cross-validation. Average prediction accuracy (%), selection robustness (%) and selected subset size are presented. EFS* denotes the proposed method. All denotes prediction in the original feature space.

Method	**Lung Tumor Data**			**Esophageal Tumor Data**		
	Accuracy	Robustness	Subset size	Accuracy	Robustness	Subset size
All	76	n/a	52	64	n/a	29
SFS	84	60	3	53	63	3
SFFS	72	54	4	**81**	53	3
SVMRFE	92	57	5	75	80	5
EFS*	**100**	**91**	4	78	**94**	3

average prediction accuracy and the selected subset size were calculated. Based on feature frequency statistics, the robustness of selection methods was evaluated using the criteria introduced in [1]. All these results are summarized on Table 3.10, in which experiments of SVM with all features are presented too as baseline for comparison. As can be seen, the proposed EFS method leads to much higher robustness of selected features. It also has the best classification performance on the lung tumor data, and the second best classification accuracy on the esophageal tumor data. On the latter dataset, EFS has led to one more misclassified patient than SFFS.

The four features robustly selected by EFS in Lung Tumour are one SUV-based feature (SUVmax during radiotherapy) and three texture-based features; while the three features robustly selected in Esophageal Tumour are one SUV-based feature (TLG before the treatment) and two clinical features.

3.3.3.2 Prediction Performance

We further tested whether feature subsets selected by EFS are applicable for other classifiers. To this end, Artificial Neural Networks (ANN), SVM, EK-NN, and the modified EK-NN (mEK-NN) discussed in Section 3.2.1 were studied. The scaled conjugate gradient back-propagation network was used here in testing ANN. The number of neurons in the hidden-layer was empirically set as 10.

In the leave-one-out cross validation (LOOCV) protocol, the selected feature subsets and all features were fed in these classifiers. The average classification accuracy is summarized

Table 3.11: Comparing the average prediction accuracy of features selected by EFS with all features using different classifiers. mEK-NN* denotes the proposed classification method.

Classifier	**Lung Tumor Data**		**Esophageal Tumor Data**	
	without EFS	with EFS	without EFS	with EFS
ANN	68	92	67	83
SVM	76	**100**	64	81
EK-NN	68	96	64	83
mEK-NN*	56	**100**	53	**89**

in Table 3.11. As can be seen, the proposed EFS improves all classifiers' prediction accuracy.

3.4 Conclusion

In this chapter, we addressed the problem of learning effectively from insufficient and uncertain data. The contribution is threefold. First, we proposed a variant of the EK-NN method based on a hybrid Dempster+Yager rule, which transfers part of the conflicting mass to the frame of discernment. This new mass construction method results in less specific mass functions than those obtained using the orignal EK-NN method introduced in [87]. The second contribution is a feature selection method that finds informative feature subsets by minimizing a special objective function using mixed integer genetic algorithm. This objective function is designed to minimize the imprecision of the mass functions, so as to obtain feature subspaces that maximize the separation between classes. Finally, the third contribution is a two-step classification strategy, which was shown to further improve classification accuracy by using already classified objects as additional pieces of evidence. These three improvements of the EK-NN method were assessed separately and jointly using several synthetic, real and clinical datasets. The proposed procedures were shown to have excellent performance as compared to other state-of-art feature selection and classification algorithms.

Dissimilarity Metric Learning in the Belief Function Framework

The Evidential K-Nearest-Neighbor (EK-NN) method provided a global treatment of imperfect knowledge regarding the class membership of training patterns. It has outperformed traditional K-NN rules in many applications, but still shares some of their basic limitations, e.g., 1) classification accuracy depends heavily on how to quantify the dissimilarity between different patterns and 2) no guarantee for satisfactory performance when training patterns contain unreliable input features. In this chapter, we propose to address these issues by learning an adaptive metric, using a low-dimensional transformation of the input space, so as to maximize both the accuracy and efficiency of the EK-NN classification. To this end, a novel loss function to learn the dissimilarity metric is constructed. It consists of two terms: the first one quantifies the imprecision regarding the class membership of each training pattern; while, by means of feature selection, the second one controls the influence of unreliable input features on the output linear transformation. The proposed method has been compared with some other metric learning methods on several synthetic and real data sets. The best performance was obtained by the proposed method.

4.1 Introduction

The K-nearest neighbor (K-NN) rule [148] is one of the most well-known pattern classification algorithms. As a case-based learning method without need of any prior assumptions [128], the K-NN classifier has been widely used in practice thanks to its simplicity. The original voting K-NN [148] assigns an object into the class represented by its majority nearest neighbors in the training set, while the information concerning the dissimilarity (distance) between the object and its neighbors is neglected. Then, the weighted K-NN [149]

has been proposed, in which this dissimilarity is imported into the classification procedure. However, in the case of uncertain and imprecise data, many samples may be corrupted with noise or located in highly overlapping areas; consequently, it becomes difficult for these classical K-NN classifiers to obtain satisfactory classification results.

To endow the K-NN method with the capability to handle uncertain information, Denœux has extended it in the belief function framework. An Evidential K-NN (EK-NN) rule has been proposed in [87], and further optimized in [88]. The EK-NN rule provides a global treatment of partial knowledge regarding the class membership of training patterns. Ambiguity and distance reject options are also taken into account based on the concepts of lower and upper expected losses [122].

The EK-NN method has outperformed other traditional K-NN methods in many situations when using the same information [88], whereas they still have some identical features: 1) the performances of the K-NN rules are strongly influenced by the chosen dissimilarity between different patterns. Better than directly using the simple Euclidean distance measure (such as in the original EK-NN), an adaptive dissimilarity metric tailored for the application should ensure better classification performance; 2) the efficiency of the K-NN rules substantially decrease when the dimensionality of the input data increases.

We propose a solution based on dissimilarity metric learning to deal with these inherent drawbacks of the K-NN classifications. Given an input space $\mathbf{X}$, the metric learning problem can be formulated as finding a transformation matrix A, such that the dissimilarity between any two patterns can be defined in the transformed space $Z = A\mathbf{X}$ [150]. Various studies have demonstrated that a properly learnt dissimilarity measure can dramatically boost the performance of the distance-based learning methods [151–156]. Even with a linear transformation of the input space [157–159], the K-NN classification can reach significant improvement. In [157], Goldberger et al. proposed a metric learning method called Neighborhood Component Analysis (NCA), which maximizes the expected leave-one-out classification accuracy from a stochastic version of the K-NN classification. Based on a softmax probability distribution defined in the transformed space, NCA labels each query instance by the majority vote of all training samples. As a main advantage of NCA, a continuous and differentiable cost function in respect of the linear transformation matrix A is deduced. This cost function can be minimized by gradient descent. The learnt matrix A can also be forced to be low-rank, thus accelerating K-NN test and facilitating class structure visualization. Although the cost used in NCA is differentiable, it seems to be

sensitive to the initialization. Inspired by NCA, Weinberger et al. proposed a Large Margin Nearest Neighbor (LMNN) method to learn a Mahalanobis distance metric for K-NN classification [159]. LMNN attempts to classify the K nearest neighbors as the same class label, under the constraint that different classes should be separated by a large margin. The learning problem is formulated as a semi-definite programming problem. The corresponding cost function consists of two terms; the first term penalizes large dissimilarities between instances with the same class label in a predefined neighborhood; while as a hinge loss, the second term penalizes small dissimilarities between instances with different class labels in the whole training pool. As a convex function in respect of the matrix A, the cost function of LMNN can be optimized efficiently.

Different from the global learning methods such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), both NCA and LMNN can adapt to the local structure of the application at hand. By learning a local dissimilarity metric, they effectively improved the K-NN classification accuracy in many situations. However, since they were not designed specifically for tackling data that contains unreliable input features, their performance can severely decline with this kind of imperfect information.

In this chapter, our goal is to maximize the accuracy and efficiency of the EK-NN classifier on data that contains unreliable input features. To this end, we propose to learn an adaptive dissimilarity metric from this kind of imperfect data in the belief function framework. By using samples in the training pool as independent items of evidence, the belief regarding the class membership of each instance is modeled and refined using DST. A specific cost function consisting of two terms is constructed for learning a low-dimensional transformation matrix A. The first term attempts to minimize the imprecision regarding the class membership of each instance. The $\ell_{2,1}$-norm regularization of A acts as the second term, considering its good property for feature selection as already shown in muti-task learning [160], multiclass classification [161], semi-supervised learning [162], etc. By means of feature selection, it aims to manage the influence of unreliable input features on the output transformation. The proposed cost function is solved efficiently by the proximal forward-backward splitting algorithm [163]. The influence of the sparsity regularization is tuned according to the application at hand. Finally, a low-dimensional transformation of the input space is realized to greatly separate instances of different classes, therefore increasing the classification accuracy and reducing the searching time of the EK-NN classifier simultaneously.

The rest of this chapter is organized as follows. The proposed metric learning method based on DST is then introduced in Section 4.2. In Section 4.3, the proposed method is tested on both synthetic and real data sets, and some comparison with other methods is presented. Finally, we conclude this chapter in Section 4.4.

4.2 Evidential Dissimilarity Metric Learning

A new approach, called evidential dissimilarity metric learning (EDML), is proposed in this section. By selecting the most informative features to learn an adaptive dissimilarity measure on training samples, EDML aims to maximize both the accuracy and efficient of the EK-NN classifier in a low-dimensional feature subspace.

4.2.1 Criterion of EDML

Let $\{(X_i, Y_i)|i = 1, \cdots, N\}$ be a collection of N training pairs, in which $X_i = [x_1, \cdots, x_V]^T$ is the ith observation with V input features, and Y_i is the corresponding class label taking values in a frame of discernment $\Omega = \{\omega_1, \cdots, \omega_c\}$. Assume the dissimilarity between instances X_i and X_j can be quantified by a squared distance measure:

$$d^2(X_i, X_j) = (X_i - X_j)^T A^T A(X_i - X_j). \tag{4.1}$$

Then, EDML attempts to find an optimal matrix $A \in \mathbf{R}^{v \times V}$ under the constraint $v \ll V$. Such a linear transformation of the input space can boost the performance of the EK-NN classifier, since important features will be selected with strong impact on calculating the distance, while the influence of unreliable features will be effectively disregarded.

To learn such a matrix A, we regard each X_i as a query instance. Then, the squared distance between X_i and X_j (i.e. $d^2(X_i, X_j)$) is used in (2.8), so as to quantify the evidence concerning the class membership of X_i that offered by training sample $(X_j, Y_j = \omega_q)$. Parameters α and γ used in (2.8) are restricted to be one for simplification.

Let Γ_q $(q = 1, \ldots, c)$ be the set of training samples (except X_i) belonging to the same class ω_q. Since the corresponding mass functions point to the same hypothesis (i.e. $Y_i = \omega_q$), they can be combined via Dempster's rule (i.e. (2.5)) to deduce a global mass function for

all training samples in Γ_q:

$$\begin{cases} m_i^{\Gamma_q}(\{\omega_q\}) & = 1 - \prod\limits_{j \in \Gamma_q} \left[1 - \exp\left\{-d(X_i, X_j)\right\}\right], \\ m_i^{\Gamma_q}(\Omega) & = 1 - m_i^{\Gamma_q}(\{\omega_q\}). \end{cases} \tag{4.2}$$

For $q = 1, \ldots, c$, the global mass function $m_i^{\Gamma_q}$ quantifies the evidence refined from the training pool that supports the assertion $Y_i = \omega_q$. The mass of belief $m_i^{\Gamma_q}(\Omega)$ measures the imprecision of this evidence. In other words, it can be regarded as the calculation of the unreliability of the hypothesis $Y_i = \omega_q$. *If the actual value of* Y_i *is* ω_q*, the corresponding imprecision should then close to zero, i.e.,* $m_i^{\Gamma_q}(\Omega) \approx 0$*; in contrast, imprecision pertaining to other hypotheses should close to one, i.e.,* $m_i^{\Gamma_r}(\Omega) \approx 1$*, for* $\forall r \neq q$. According to this assumption, we propose to represent the prediction loss for training sample (X_i, Y_i) as

$$loss_i(A) = \sum_{q=1}^{c} t_{i,q} \cdot \left\{ 1 - m_i^{\Gamma_q}(\{\omega_q\}) \cdot \prod_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega) \right\}^2, \tag{4.3}$$

where $t_{i,q}$ is the qth element of a binary vector $t_i = \{t_{i,1}, \ldots, t_{i,c}\}$, with $t_{i,q} = 1$ if and only if $Y_i = \omega_q$. When $Y_i = \omega_q$ is true, minimizing $loss_i(A)$ can force both $m_i^{\Gamma_q}(\{\omega_q\}) = 1 - m_i^{\Gamma_q}(\Omega)$ and $\prod\limits_{r \neq q}^{c} m_i^{\Gamma_r}(\Omega)$ to approach one as far as possible, thus achieving the goal to maximize the reliability of the right hypothesis ($Y_i = \omega_q$) but to minimize the reliability of other assertions. As the result, the learnt matrix A can lead X_i only close to samples from the same class in the transformed space, thus protecting the classification performance of the EK-NN method.

Therefore, for all training samples, the loss function in respect of the transformation matrix A can be finally defined as

$$l(A) = \frac{1}{N} \sum_{i=1}^{N} loss_i(A) + \lambda ||A||_{2,1}, \tag{4.4}$$

where $loss_i(A)$ represents the learning cost for training sample (X_i, Y_i) that quantified by (4.3). The $\ell_{2,1}$-norm sparsity regularization defined as

$$||A||_{2,1} = \sum_{j=1}^{V} \left(\sum_{i=1}^{v} A_{i,j}^2 \right)^{1/2} \tag{4.5}$$

is imported to select input features. By forcing columns of the transformation matrix A to be zero during the learning procedure, this sparsity term only selects the most reliable input features to calculate the linear transformation, thus controlling the influence of unreliable

input features on the output low-dimensional transformed space. Scalar λ is a hyper-parameter that controls the influence of this regularization. Generally speaking, a too small λ may fail to limit the influence of unreliable input features; while, a too large λ may also delete informative features.

4.2.2 Optimization

Since $loss_i$ (4.3) is differentiable in respect of matrix A while $||A||_{2,1}$ (4.5) is partly smooth (it is non-smooth when and only when $A = 0$), the proximal Forward-Backward splitting (FBS) algorithms [163, 164], which belong to the class of first order methods, are efficient alternatives to solve the proposed loss function (4.4). More specifically, as an improved version of the classical FBS methods, the Beck-Teboulle proximal gradient algorithm [165] is used considering its computational simplicity and fast convergence rate.

In general, each iteration of the FBS algorithms can be broken up into a gradient descent step using $\frac{1}{N}\sum_{i=1}^{N} loss_i(A)$, followed by a proximal operation using $||A||_{2,1}$. According to (4.1)-(4.3), the derivative of $loss_i$ concerning A (i.e. $\partial loss_i/\partial A$) can be deduced as

$$\begin{aligned}\frac{\partial loss_i}{\partial A} =& \sum_{q=1}^{c} 2t_{i,q}\left\{1 - m_i^{\Gamma_q}(\{\omega_q\})\prod_{r\neq q}^{c} m_i^{\Gamma_r}(\Omega)\right\} \\ & \left\{ - \frac{\partial m_i^{\Gamma_q}(\{\omega_q\})}{\partial A}\prod_{r\neq q}^{c} m_i^{\Gamma_r}(\Omega)\right. \\ & \left. - m_i^{\Gamma_q}(\{\omega_q\})\sum_{r\neq q}^{c}\frac{\partial m_i^{\Gamma_r}(\Omega)}{\partial A}\prod_{s\neq r,q}^{c} m_i^{\Gamma_s}(\Omega)\right\}.\end{aligned} \tag{4.6}$$

In which, value $m_i^{\Gamma_q}$ is calculated via (4.2), and for $\forall q = 1, \ldots, c$,

$$\frac{\partial m_i^{\Gamma_q}(\{\omega_q\})}{\partial A} = -\sum_{j\in\Gamma_q}\frac{\partial m_{ij}(\Omega)}{\partial A}\prod_{l\in\Gamma_q\backslash j} m_{il}(\Omega); \tag{4.7}$$

While, mass m_{ij} is determined using (2.8) and (4.1), and

$$\frac{\partial m_{ij}(\Omega)}{\partial A} = 2m_{ij}(\{\omega_q\})A(X_i - X_j)(X_i - X_j)^T. \tag{4.8}$$

Based on (4.6)-(4.8), the Beck-Teboulle proximal gradient algorithm executes as the form shown in Algorithm 1, so as to deduce an optimal or at least sub-optimal low-dimensional transformation matrix A. To facilitate the optimization procedure, classical metric learning methods (e.g. PCA) can be used to generate the initialization (i.e. $A^{(0)}$) for the proposed method. The learnt matrix A is then applied in (4.1) to measure the dissimilarity between different instances, and finally used in the EK-NN classification.

Algorithm 1: Beck-Teboulle proximal gradient algorithm [165].

Initialize $A^{(0)} \in \mathbf{R}^{v \times V}$ and $\beta > 0$, set $H^{(0)} = A^{(0)}$ and $t^{(0)} = 1$;

for $n = 0, 1, 2, \ldots, n_{max}$ **do**

$$G^{(n)} = H^{(n)} - \frac{\beta^{-1}}{N} \sum_{i=1}^{N} \frac{\partial loss_i}{\partial A}|_{A=H^{(n)}} \ ;$$

$$A^{(n+1)} = \arg\min_{B} \left\{ \lambda ||B||_{2,1} + \frac{\beta}{2} ||B - G^{(n)}||^2 \right\} \ ;$$

$$t^{(n+1)} = [\sqrt{4t^{(n)^2} + 1} + 1]/2 \ ;$$

$$\delta^{(n)} = 1 + [t^{(n)} - 1]/t^{(n+1)};$$

$$H^{(n+1)} = A^{(n)} + \delta^{(n)}[A^{(n+1)} - A^{(n)}] \ ;$$

4.3 Experimental Results

The presented experiments consist of five parts. In the first part, the proposed method, namely EDML, was evaluated on a synthetic data set. The proportion of unreliable (noisy and imprecise) features in this synthetic data was varied to assess the robustness of EDML under different situations. In the second part, EDML was evaluated on several real data sets. The corresponding classification accuracy was compared with some other metric learning methods. The parameters used in the proposed method were also studied. In the third part, we studied the parameters of EDML. In the forth part, we further compared the two-dimensional visualization performance of different metric learning methods, so as to evaluate whether the proposed method can effectively separate instances from different classes in low-dimensional subspaces. Finally, the proposed method was applied to predict cancer therapy outcomes, and the same clinical data used in Chapter 3.3.3 were adopted to evaluate its performance.

4.3.1 Performance on Synthetic Data

The studied synthetic data sets were generated using a process similar to the one described in Chapter 3.3.1.1. The feature space contains n_r relevant features, n_u irrelevant (noisy) features, and also n_i imprecise features copied as the cubic of the relevant features. The numbers of relevant, irrelevant and imprecise features were set, respectively, as $n_r = 2$, $n_u \in \{6, 16, 26, 36, 46\}$ and $n_i = 2$ to simulate five different situations. Under each situation, we generated 150 training instances and the same number of testing instances. PCA, NCA, LMNN and the proposed EDML methods were executed to learn a two-dimensional

Table 4.1: Classification accuracy (both training and testing, in %) of the EK-NN based on different metric learning methods. In the studied synthetic data sets, $n_r = 2$ and $n_i = 2$. EDML-FS and EDML denote, respectively, the proposed method with/without the $\ell_{2,1}$-norm sparsity regularization. Performance of the SVM and ENN classifiers joint with PCA were also presented as two baselines for comparison.

	n_u	SVM	ENN	PCA	NCA	LMNN	EDML	EDML-FS
training	6	92.00	91.33	90.67	99.33	98.00	98.67	99.33
	16	83.33	85.33	84.67	100.00	96.00	99.33	100.00
	26	81.33	83.33	74.00	100.00	96.67	100.00	100.00
	36	77.33	76.00	76.00	100.00	100.00	99.33	100.00
	46	76.67	74.67	68.67	100.00	99.33	99.33	100.00
testing	6	85.33	84.00	84.00	91.33	90.00	86.00	**94.67**
	16	74.00	72.00	73.33	84.00	86.67	86.00	**92.00**
	26	66.67	69.33	64.67	78.67	78.67	84.67	**90.00**
	36	69.33	69.67	62.00	70.00	78.00	76.67	**95.33**
	46	64.67	66.00	57.33	82.67	76.67	76.67	**94.00**

dissimilarity metric A (i.e. $\in \mathbf{R}^{n_r \times (n_r+n_u+n_i)}$) on the training set. The obtained metric A was then used in the EK-NN to classify both the training and testing samples. As two baselines, results obtained by the SVM and Evidential Neural Network (ENN) [89] classifiers joint with PCA were also included for comparison.

Parameters of each method used in this experiment (four metric learning methods, i.e., PCA, NCA, LMNN and EDML, and three classifiers, i.e., EK-NN, ENN and SVM) can be summarized as follows:

- For LMNN, as suggested by [159], parameters K and μ were set as $K = 3$ and $\mu = 0.5$.
- For the proposed EDML, a rough grid search strategy was used to select an appropriate λ from $\{0.005, 0.007, 0.009\}$ according to the training performance. More specifically, the EK-NN classifier was adopted to classify training data using learnt metric that obtained by each possible λ; then, the optional λ that led to the highest classification accuracy was selected.

Table 4.2: Properties of the five real data sets studied in Section 4.3.2.

data sets	classes	input features	instances
Wine	3	13	178
Seeds	3	7	210
Soybean	4	35	47
LSVT	2	309	126
Faces	40	100	400

- For the EK-NN classifier, parameters α and γ were optimized via the operation proposed in [88]. The number of nearest neighbors was set as $K = 3$.
- For the SVM, the gaussian kernel was used with the radial basis $\sigma = 1$.
- For the ENN classifier, the number of prototypes per class was set as 5.

It is worth illustrating that the parameters of the compared methods were always kept the same in the sequel experiments.

Finally, the training and testing (more important) accuracy (in %) obtained by different metric learning methods are summarized in Table 4.1, in which EDML (manually set $\lambda = 0$) and EDML-FS (namely EDML joint with Feature Selection) represent, respectively, the proposed method without/with the sparsity regularization. As can be seen, the proposed EDML-FS led to higher testing accuracy than other methods under all the five different situations. It is also worth noting that the difference increased following the augment of unreliable input features, which reveals that the proposed method is stable and immune to severely deteriorated input information.

Table 4.3: The best training and the corresponding testing accuracy (ave±std, in %) obtained by different methods with $v \in \{2, 3, \ldots, 15\}$. EDML-FS and EDML denote, respectively, the proposed method with/without the sparsity regularization. Results of the SVM and ENN joint with PCA were also presented as two baselines for comparison.

		SVM	**ENN**	**PCA**	**NCA**	**LMNN**	**EDML**	**EDML-FS**
training	Wine	99.95±0.19	100.00±0.00	97.65±1.24	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	Seeds	99.58±0.47	98.74±0.644	92.90±1.87	97.04±1.33	95.86±1.50	99.13±0.70	98.52±1.27
	Soybean	100.00±0.00	100.00±0.00	98.85±1.72	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	LSVT	99.52±0.83	93.82±1.28	80.71±3.94	98.63±1.69	100.00±0.00	96.45±1.46	97.14±1.33
	Faces	83.43±2.69	99.99±0.05	90.28±1.34	99.87±0.24	100.00±0.00	99.37±0.37	99.71±0.26
testing	Wine	96.15±2.50	97.51±2.30	95.90±2.63	96.97±2.23	97.88±1.68	96.46±2.95	**97.98±1.86**
	Seeds	91.78±2.93	92.92±2.96	92.25±3.14	93.85±2.88	94.54±2.78	94.48±2.45	**95.49±1.84**
	Soybean	**100.00±0.00**	98.00±3.83	98.28±3.97	99.86±1.00	99.44±2.80	**100.00±0.00**	**100.00±0.00**
	LSVT	80.42±4.82	85.21±3.94	80.92±5.76	82.57±6.35	82.13±5.40	85.03±4.98	**86.09±4.63**
	Faces	65.88±4.05	85.23±3.08	89.48±2.20	89.18±3.50	**97.63±1.66**	93.40±2.13	97.08±1.55

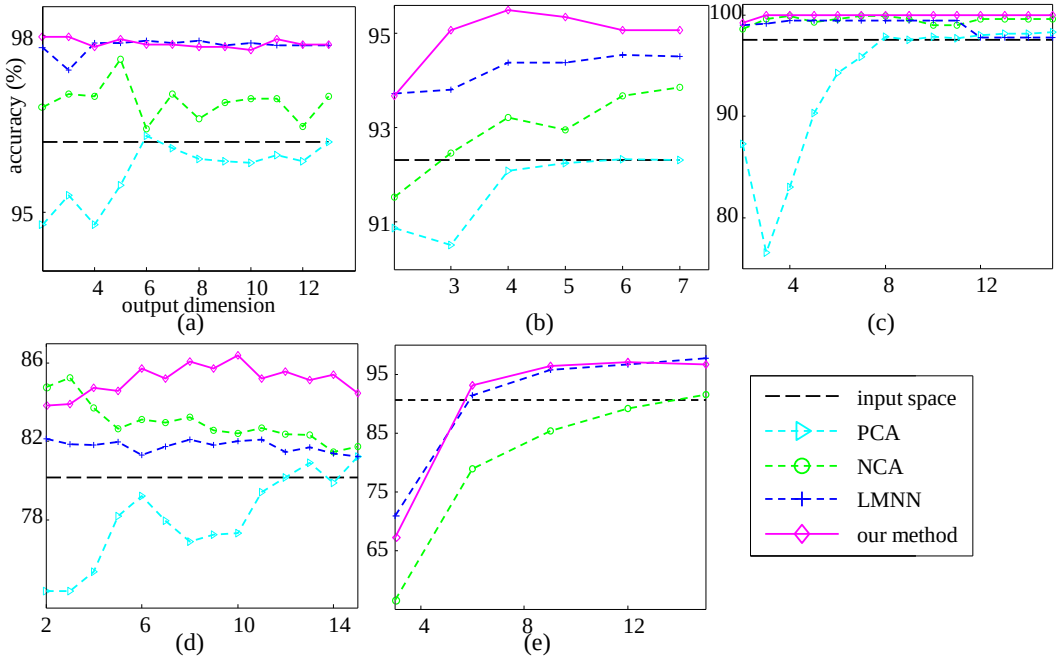


Figure 4.1: Average testing accuracy obtained by different metric learning methods: (a) Wine data, (b) Seeds data, (c) Soybean-small data, (d) LSVT data and (e) Faces data. In each subfigure, the horizontal axis represents the output dimension (i.e. v) of the learnt transformation A, while the vertical axis represents the corresponding classification accuracy (in %).

4.3.2 Performance on Real Data

The proposed method was further evaluated using five real data sets of varying input features and classes. Four of these data sets (Wine, Seeds, Soybean-small and LSVT voice rehabilitation [166] data) were downloaded from the UCI Machine Learning Repository[1]. The other one is the Olivetti face recognition data set[2]. As a preprocessing operation for the Face data, we down-sampled the images to 38×31 pixels and used PCA to further reduce the dimensionality to 100. Properties of all the five data sets are briefly summarized in Table 4.2.

The training and testing instances were randomly generated with 70/30 splitting, and repeated 50 times. Under each random split, we used PCA, NCA, LMNN and the proposed EDML methods, respectively, to learn a low-dimensional dissimilarity metric A (i.e. $\in \mathbf{R}^{v \times V}$ with $v \leq V$) on the training data; then used it in the EK-NN to classify both the training and testing instances. Parameters of compared methods were the same as that

[1]Please see at https://archive.ics.uci.edu/ml/index.html

[2]Please see at http://www.uk.research.att.com/facedatabase.html.

used in the last experiment (namely Section 4.3.1). For the proposed method, the hyper-parameter λ was still determined by a rough grid search strategy according to the training performance. More specifically, the EK-NN classifier was adopted to classify training data using learnt dissimilarity metric that corresponds to each optional λ; then, parameter λ that led to the highest classification accuracy was used. On average, good results were obtained with λ between $[0.0005, 0.01]$ for the five data studied in this experiment.

The value of the output dimension v was orderly set as $\{2, 3, \ldots, 15\}$. Then, the average testing accuracy with different v was calculated and is shown in Figure 4.1. As can be seen, the proposed method consistently performed well on these data sets as compared with other methods. More specifically, LMNN (blue line) and EDML (magenta line) had comparable testing accuracy on Wine and Faces data sets; NCA (green line), LMNN and EDML resulted in almost the same performance (EDML was slightly better) on the soybean-small data set; and EDML yielded the best performance on the other two data (Seeds and LSVT).

To further analyze the experimental results obtained on these real data sets, we computed the average training performance as a criterion to select the best output dimension v (from $\{2, 3, \ldots, 15\}$) for the learnt dissimilarity metric A. The best training accuracy and the corresponding testing accuracy (more important) for each method are summarized in Table 4.3, in which results obtained by the SVM and ENN joint with PCA are also presented as two baselines for comparison. As in the former subsection, EDML (manually set $\lambda = 0$) and EDML-FS represented the proposed method without/with the sparsity regularization. From Table 4.3, it can be found that EDML-FS consistently yielded better performance than other methods on the first four data sets, especially on the LSVT data. This is mainly because the proposed method only selected the most informative features (from all the three hundred input features) to calculate the linear transformation. LMNN slightly outperformed our method on the Face data set, but with a slight higher variance. In addition, we can also see that, thanks to the $\ell_{2,1}$-norm sparsity regularization, EDML-FS performed better than EDML.

4.3.3 Parameter Analysis

4.3.3.1 Output Dimension

As discussed above, the best output dimension v (from $\{2, 3, \ldots, 15\}$) for the five real data sets studied in the last subsection was determined according to the training performance.

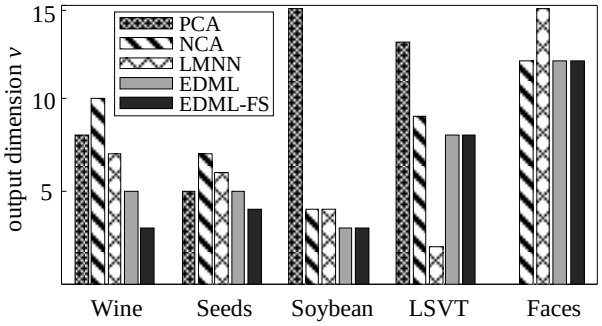


Figure 4.2: The best output dimension v (between 2 and 15) according to the training performance obtained by different methods on the five real data sets.

Therefore, besides the classification accuracy presented in Table 4.3, the corresponding output dimension obtained by different methods on these real data sets was also summarized and is shown in Figure 4.2.

4.3.3.2 Regularization Parameter

The tuning parameter λ in the loss function (4.5) controls the effect of the sparsity regularization on the output low-dimensional transformation. It should be tuned specifically for each data set at hand. Generally speaking, a too small λ may fails to limit the influence of unreliable input features, while a too large λ may also removes many significant input features. On average, good results were obtained with λ between $[0.0005, 0.01]$ for all the five data studied in the last subsection.

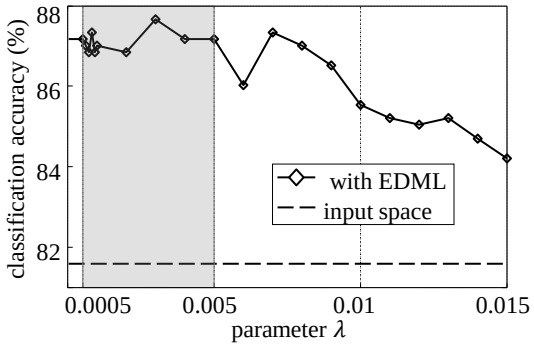


Figure 4.3: Average testing accuracy on the LSVT data set with regard to the hyper-parameter λ. The output dimension was set as $v = 5$. The dashed line represents the accuracy obtained in the input space.

In this experiment, the LSVT data set was used as an example to further analysis the

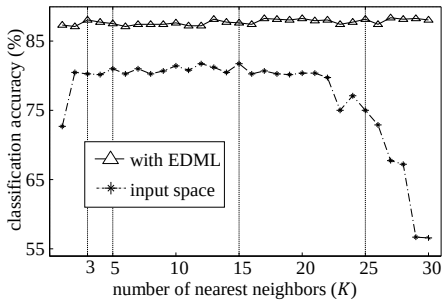


Figure 4.4: Average accuracy of the EK-NN classification on the LSVT data set with regard to the number of nearest neighbors K. The output dimension was set as $v = 5$.

influence of the parameter λ. The training and testing data were generated with 70/30 splitting, and repeated 20 times. We orderly selected a λ from $\{0, 0.0005, 0.0006, \ldots, 0.001, 0.002, \ldots, 0.015\}$ to learn a low-dimensional transformation (with $v = 5$) of the input space. Then, the EK-NN classifier (with $K = 3$) was used to classify the testing instances on the transformed space. For all the 20 random splits, the average testing accuracy (in %) with regard to λ is finally shown in Figure 4.3, in which the horizontal line represents the average accuracy of the EK-NN classification in the input space. As can be seen, relatively high performance on this data set is obtained with λ between $[0.0005, 0.01]$. The classification is less sensitive in the region $[0.0005, 0.005]$ than in other regions of λ.

4.3.3.3 Number of Nearest Neighbors

We also studied the parameter K of the EK-NN classification with the dissimilarity metric learnt by the proposed method. Still on the LSVT data set, the training and testing data were generated with 70/30 splitting, and repeated 20 times. Under each random split, we used the proposed method to learn a low-dimensional transformation of the input space. The output dimension and the regularization parameter were set as $v = 5$ and $\lambda = 0.002$. Then, the EK-NN classifier with $K = \{1, 2, \ldots, 30\}$ was orderly executed to classify the testing instances in the transformed space. As for comparison, the EK-NN classifier with the same K was also directly executed in the input space to classify the testing instances. The average testing accuracy with regard to K is finally summarized in Figure 4.4. It can be found that, with a metric learnt by the proposed method, the EK-NN classification always has higher accuracy on this data set than directly using the Euclidian distance in the input space. In addition, we can also see that the proposed method is robust to the parameter K.

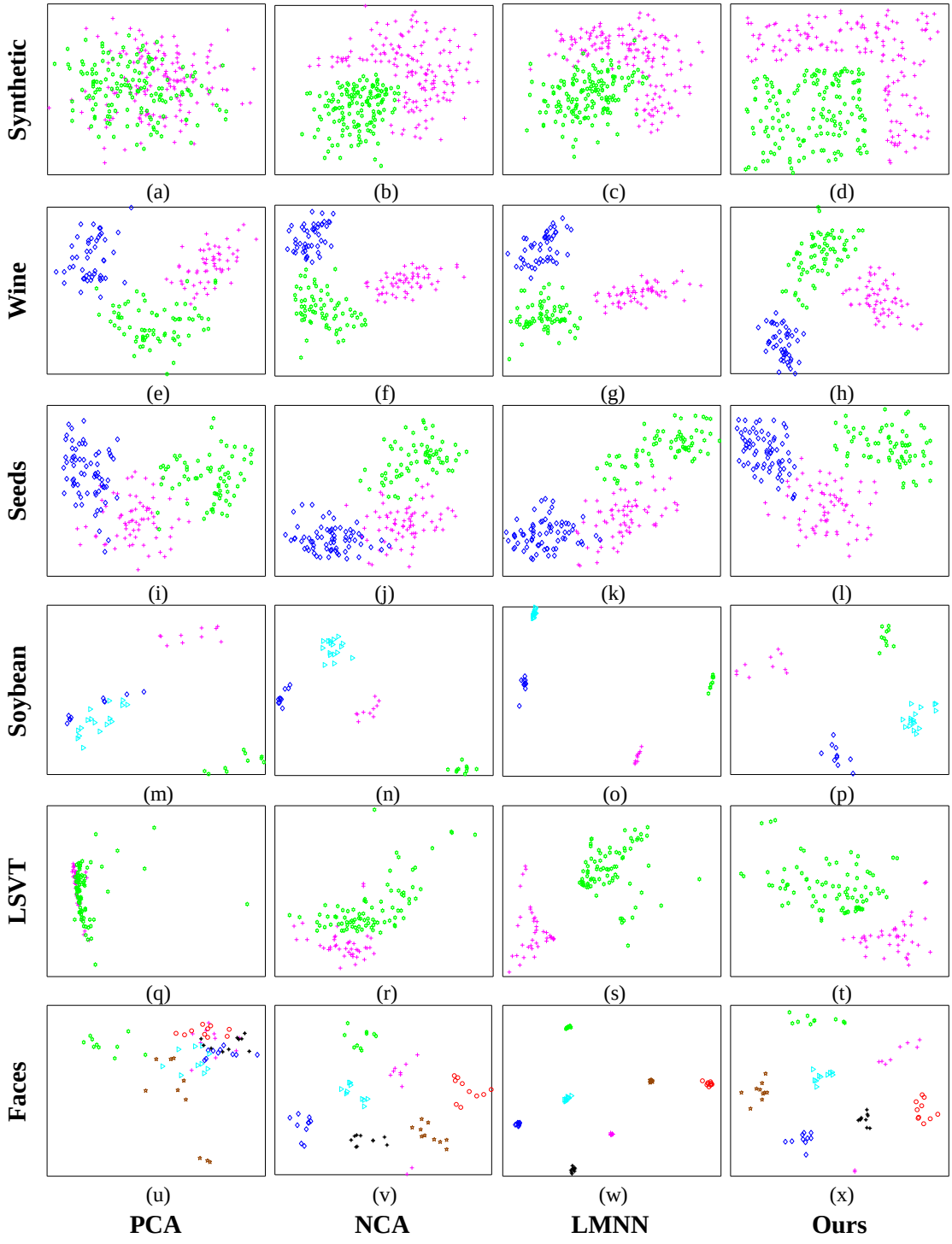


Figure 4.5: Two-dimensional transformation results obtained by PCA, NCA, LMNN and the proposed method (orderly from the first to the forth column). (a)-(d) on synthetic data; (e)-(h) on Wine data; (i)-(l) on Seeds data; (m)-(p) on Soybean data; (q)-(t) on LSVT data; (u)-(x) on Faces data;

4.3.4 Two-Dimensional Visualization

To further evaluate whether the proposed method can effectively separate instances from different classes in low-dimensional transformation space, we visualized the dimension reduction in 2D, as shown in Figure 4.5. PCA, NCA, LMNN and the proposed method were still compared on one synthetic and five real data sets used in the previous subsections. The input feature space for the synthetic data was set as fifty ($n_r = 2$,$n_i = 2$ and $n_u = 46$). For simplicity, only the first seven classes were studied in the Faces data. From the obtained results we can see that instances from different classes were always well separated by our method on all the six data sets. It led to the largest margin on the synthetic data, and the most satisfying separation on the Seeds data. In contrast, NCA did not separate the LSVT data perfectly; while LMNN resulted in large overlaps on the synthetic data.

4.3.5 Performance on Clinical Data

Using the two clinical data sets that have been studied in Chapter 3.3.3, we compared the proposed EDML method with several feature transformation methods, namely PCA, linear discriminant analysis (LDA), NCA and kernel PCA (K-PCA) [167]; and several feature selection methods, namely T-test, Information Gain (IG), Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) [139]. Features of the above datasets are described in detail in Chapter 5.

The leave-one-out cross-validation (LOOCV) procedure was used for evaluation. For other feature selection or transformation methods (except NCA, since it was designed specifically for the K-NN classifers), after learning a low-dimensional subspace, the SVM (Gaussian kernel with $\sigma = 1$ was empirically chosen) classifier was used to predict class labels of both training instances and the left testing instance; while the EK-NN classifier (K was empirically set as 3) was used with NCA and the proposed method. Tuning parameter λ for EDML-FS was determined by a rough grid search strategy. The EK-NN classifier was adopted to classify training data using learnt metric that obtained by each optional λ; then, the optional λ that led to the highest classification accuracy was selected. The dimension of output subspace was chosen between two to five according to the minimum average testing error. Finally, the average training and testing accuracy for all methods are summarized in Table 4.4, in which results obtained by the SVM and EK-NN in the input space, and by our method without feature selection (namely with $\lambda = 0$) are also presented

Table 4.4: Comparing prediction accuracy (in %) of different methods. EDML-FS* and EDML* denote, respectively, the proposed method with/without the $\ell_{2,1}$-norm sparse regularization.

Method	**Lung Tumor Data**		**Esophageal Tumor Data**	
	training	testing	training	testing
EK-NN	69.50	60.00	63.73	61.11
SVM	100.00	76.00	100.00	63.89
T-test	99.67	72.00	75.56	66.67
IG	86.50	68.00	88.57	75.00
SFS	95.67	84.00	85.63	52.78
SFFS	64.33	72.00	59.68	80.56
PCA	88.33	80.00	59.60	55.56
LDA	100.00	52.00	100.00	55.56
NCA	99.50	80.00	94.21	69.44
K-PCA	81.33	80.00	71.19	72.22
EDML*	95.83	**88.00**	88.02	63.89
EDML-FS*	100.00	**88.00**	97.46	**83.33**

for comparison. It can be observed that the proposed method, especially EDML-FS, leads to higher testing performance than other methods. Although LDA results in larger training accuracy than our method, the worst testing performance is obtained. It maybe because the studied data sets were too small, therefore the covariance matrix obtained by LDA has been badly scaled. It is also worth noting that EDML and EDML-FS have the same testing performance on the lung tumor data, while EDML-FS performs much better on the esophageal tumor data than EDML. This result maybe can be explained from two different aspects: firstly, the lung tumor data is easier to be separated than the esophageal tumor data, hence the difference became small; on the other hand, it perhaps also demonstrates that the sparse term can play a real role to improve the prediction under complex situation, such as on the esophageal tumor data.

Furthermore, we visualized the dimension reduction in 2D achieved using PCA, NCA, EDML and EDML-FS methods, as shown in Figure 4.6. It can be seen that different classes in both data sets are better separated by our methods than using other methods. The best

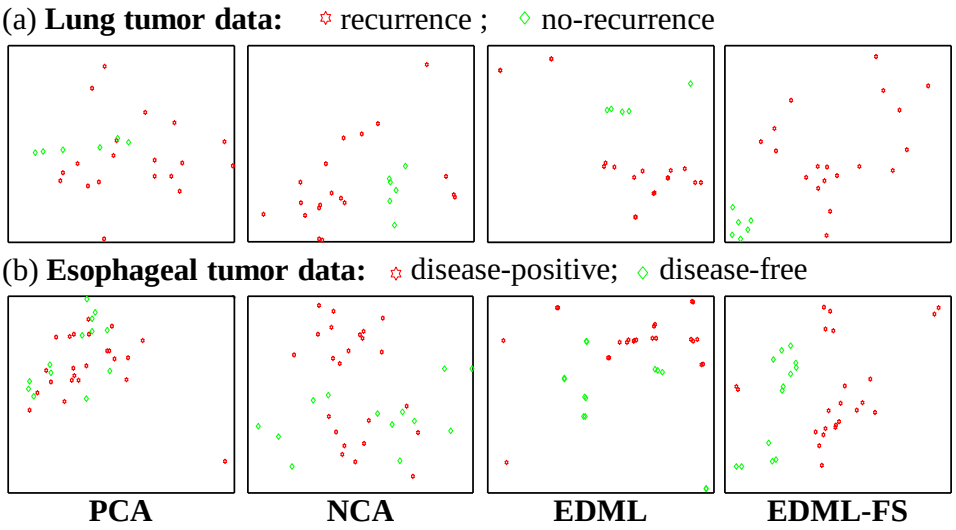


Figure 4.6: Two-dimensional transformation results of PCA, NCA, our EDML (without feature selection, i.e., $\lambda = 0$) and EDML-FS.

separation is achieved using our method with feature selection (EDML-FS).

4.4 Conclusion

To optimize the performance of the EK-NN classification on imperfect data sets, an approach based on belief functions has been proposed to learn a dissimilarity metric specifying for the application at hand. By treating other training patterns as different sources of information, the belief concerning the class membership of each query pattern has been quantified and refined in the belief function framework. A specific loss function consisting of two terms has been developed for metric learning under uncertainty, in which the first term is used to minimize the imprecision regarding each instance's class membership, while the second term is the $\ell_{2,1}$-norm sparsity regularization of the low-dimensional transformation matrix. Through a feature selection procedure, it serves to limit the influence of uncertainty and/or imprecise input features. The proposed method has been evaluated on both synthetic and real datasets, consistently showing good performance with regard to classification accuracy, computational efficiency, class structure visualization. Moreover, it has also proved that the proposed method is not sensitive to the parameter K. Experimental results obtained on two clinical data sets have also shown that the proposed method performs well in cancer therapy outcome prediction.

Robust Cancer Treatment Outcome Prediction Dealing with Small-Sized and Imbalanced Data from FDG-PET Images

In this chapter, we propose a prediction system primarily using radiomic features extracted from FDG-PET images. Other sources of information (e.g. clinical characteristics, and genomic expressions, etc) are also gathered as the complementary knowledge for more reliable treatment outcome prediction.

The proposed system includes a feature selection method, which focuses on dealing with small-sized and imbalanced learning problem (a typical problem of clinical data) in robustly selecting discriminant feature subset for accurate outcome prediction. To this end, a specific data rebalancing procedure and specified prior knowledge are taken into account. Finally, the Evidential K-NN (EK-NN) classifier is used with selected features to output prediction results. Our prediction system has been evaluated by synthetic and clinical datasets, consistently showing good performance.

5.1 Introduction

Although the quantification of radiomic features from FDG-PET images, as well as the calculation of their temporal changes during the treatment, have been claimed to have the discriminative power [48], the solid application is still hampered by some practical difficulties:

First, *uncertainty and inaccuracy of extracted radiomic features* caused by noise and

limited resolution of imaging systems, by the effect of small tumour volumes, and also by the lack of a priori knowledge with respect to the most discriminant features.

Second, *small-sized dataset* often encountered in the medical domain, which results in a high risk of over-fitting with a relatively high-dimensional feature space.

Third, *skewed dataset* where the number of training samples from different classes are severely imbalanced, thus usually leading to poor performance for classifying the minority class.

Feature selection is a feasible solution for above challenges. It aims to select a subset of features that can facilitate data interpretation and improve prediction accuracy [134]. Univariate selection and multivariate selection are two rough categories of feature selection algorithms. According to chosen statistical measures, univariate methods utilize variable ranking as the principal selection mechanism. RELIEF (RELevance In Estimating Features) [168] is considered as one of the most successful univariate selection methods, in which a margin-based criterion is used to rank the features. FAST (Feature Assessment by Sliding Thresholds) [169], another feature ranking method, has the ability to tackle small sample size and imbalanced data problems. These univariate algorithms are simple and scalable; however, they may produce sub-optimal subsets as they ignore the interaction between features [134].

Different from ranking features, multivariate methods evaluate a subset of features ensemble. Sequential Forward Selection (SFS) and Sequential Forward Floating Selection (SFFS) [139] are two classical subset selection methods. According to the prediction accuracy of a specific classifier, and starting from an empty set, SFS repeatedly selects the best feature among the remaining features to yield a nested feature subset. Since former included features can not be deleted anymore, it has the possibility to be trapped in local minima. SFFS has been used with learning methods to automatically detect lung nodules in thoracic CT [170]. It in some sense reduces the nesting problem of SFS, but still has the risk to be sub-optimal with limited learning instances [22]. To improve the performance of forward selection methods (such as SFS and SFFS) on small-sized datasets, a Hierarchical Forward Selection (HFS) method with an advanced searching strategy was proposed by [22]. Different with SFS, HFS retains all candidate feature subsets that improve the classification accuracy in each iteration. As the result, it is more likely to obtain the most discriminative feature subset, while with the cost of increased searching time. Based on a generalization of the Support Vector Machine (SVM), Guyon et al.

embedded a Recursive Feature Elimination procedure into the construction of the SVM classifier (namely SVMRFE) [141]. The variants of this method have been successfully applied for prostate cancer volume estimation [171] and deformable registration in medical imaging [172]. Starting with all input features, and before reaching a predefined number of remaining features, SVMRFE progressively eliminates the least relevant features. It yields nested feature subsets, and has the risk of removing useful features that are complementary to others. Kernel Class Separability (KCS)-based feature selection method ranks feature subsets according to the class separability [173]. As a robust method, KCS has found promising application for tumor delineation in multi-spectral MRI images [174]. But just like univariate methods, a threshold should be manually specified for KCS to output a feature subset.

Apart from the prediction accuracy, the stability of feature selection is also an important issue. As pointed by [1], the stability of a feature selection algorithm, referring to its robustness against changing conditions (e.g., perturbations of training data), can directly effect the reliability of a learning system. A key issue of the conventional feature selection methods discussed above is the difficulty to ensure robust selection performance with severely imperfect knowledge, such as seriously imbalanced training set, and high overlapping or noisy training set.

To learn efficiently from noisy and high overlapping training dataset, a robust subset selection method, called Evidential Feature Selection (EFS), has been proposed in Chapter 3. This method allows to quantify the uncertainty and imprecision resulted by different feature subsets. A specific loss function with a sparsity constraint is minimized to find a required subset that leads to both high classification accuracy and small overlaps between different classes. Due to system noise and low-resolution of PET imaging, as well as the effect of small tumor volumes [54], in our application, the training set used for constructing the prediction system may contain imprecise or inaccurate observations. Under this condition, EFS can provide better performance than other conventional methods [175]. However, the imbalanced learning problem in feature selection (another important issue of medical data) is still left unsolved for this method.

In this chapter, we propose a new framework based on our previous work (EFS) for PET imaging based treatment outcome prediction. To this end, a data balancing procedure is added to EFS, so as to control the influence of imbalanced learning data on feature selection. In addition, to cope with small-sized datasets and to improve the subset robustness, prior

knowledge is included in EFS to guide the feature selection procedure. The loss function used in the original EFS is also changed to reduce the complexity of the prediction system. Finally, the Evidential K-NN (EK-NN) rule [87] is used with selected feature subsets to output prediction results.

The rest of this chapter is organized as follows. An improved EFS with prior knowledge and data balancing is introduced in Section 5.2. The proposed method is evaluated by three clinical datasets described in Section 5.3, and the experimental results are summarized in Section 5.4. The conclusions are presented in Section 5.5.

5.2 Method

The proposed prediction system is learnt on a dataset $\{(X_i, Y_i)|i = 1, \cdots, N\}$ of N tumor patients with already known treatment outcomes. For each patient i, vector $X_i = [x_{i,1}, \cdots, x_{i,V}]^T$ consists of V input features extracted from different sources of information. Correspondingly, label Y_i denotes the (binary) outcome after treatment. In our applications, the treatment outcomes always only have two possible values (e.g., recurrence or no-recurrence). Hence, without loss of generality, the frame of discernment (possible classes) is defined as $\Omega = \{\omega_1, \omega_2\}$ to indicate that only the binary classification problems are considered in this method.

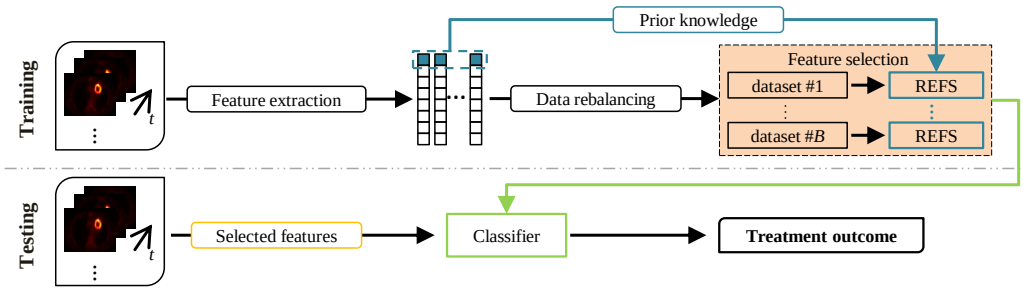


Figure 5.1: Framework of the prediction system.

5.2.1 Main Framework

The rough protocol of the prediction system is shown in Figure 5.1. To begin with, features are extracted from multi-sources of information, which include FDG-PET images of the patients acquired before and during the treatment, clinical characteristics and genomic expressions, etc. A data balancing method is then used to balance the training samples,

which are originated from two different classes, for feature selection. An improved EFS is executed to select features from the balanced datasets. During this procedure, prior knowledge is incorporated into EFS, so as to improve the robustness of the selected features. Finally, based on the selected feature subset, the Evidential K-Nearest-Neighbor (EK-NN) classification rule is trained with the original training dataset to predict the cancer treatment outcome.

Table 5.1: Definition of SUV-based features. Variable X represents SUVs in the ROI. Function $T[\cdot]$ is a binary indicator. It equals to 1 iff the argument is true. Function f maps X to L ={tumor,non-tumor} according to the threshold 40%SUV$_{max}$. Operation $|\cdot|$ calculates the number of voxels within a region.

Feature	Calculation	Description
SUV$_{max}$	$\alpha = \max(X)$	Maximum uptake in the ROI
SUV$_{mean}$	$\mu = \text{mean}(X)$	Average uptake in the ROI
SUV$_{peak}$	$\mu_\alpha = \frac{1}{\lvert N_\alpha \rvert} \sum_{x \in N_\alpha} x$	Average uptake in the neighborhood ($3 \times 3 \times 3$) of the SUV$_{max}$
MTV	$\tau = \text{sum}(T[f(X)])$	Metabolic tumor volume
TLG	$\nu = \mu \times \tau$	Total lesion glycolysis

5.2.2 Feature Extraction

To extract features, FDG-PET images for the same patient acquired at different time points are registered to the baseline image (i.e., image at initial staging) with a rigid registration method. The registration result is manually adjusted by physicians to avoid obvious misregistration. The ROIs around tumors are delineated by a relative threshold method, or manually delineated by experienced physicians when the result obtained by the threshold method is not reliable. It is worth to mention that the reproducibility of the manual tumor delineation has been evaluated in some clinical studies [41]. Three types of PET imaging features are quantified, namely SUV-based features, texture features, and the temporal changes of these two types of features.

5.2.2.1 SUV-based features

Five types of SUV-based features are calculated from the ROI of each PET stack, namely SUV_{min}, SUV_{max}, SUV_{peak}, MTV and TLG. The detail description of these features, and the formulas for calculating them are shown in Table 5.1.

5.2.2.2 Texture features

To characterize tumor uptake heterogeneity, texture features are also considered in our prediction system. As has been claimed to be effective in PET image characterization [20], Gray Level Size Zone Matrix (GLSZM) [2] is used to extract texture features. To this end, we resample voxel intensities inside the ROI to 2^3 different values. By defining the connected voxels with the same gray level as a zone, a matrix with 2^3 rows is then deduced, in which the element at row r and column s stores the number of zone with gray level r and size s. The number of columns of this matrix is determined by the size of the largest zone. Therefore, a wide and flat matrix indicates that the texture information is homogeneous in the predefined ROI; while heterogeneity when the matrix is narrow. Based on this matrix, we compute eleven variables to describe the regional heterogeneity. The formulas for calculating these GLSZM-based features are presented in Table 5.2.

5.2.2.3 Temporal changes of image features

Considering that the temporal changes of these SUV-based and GLSZM-based features may also provide discriminative value, we propose to calculate their relative difference between the baseline and the follow-up PET acquisitions as additional features. The relative difference can be generally represented as $\Delta f = (f_t - f_0)/f_0$, where f_0 and f_t denote the same kind of feature extracted from the baseline and the follow-up images, respectively.

5.2.2.4 Other features

Apart from image features, variables extracted from other sources of information may be also important knowledge that can be taken into account. Hence, patients' clinical characteristics and genomic expressions are also included in our prediction system as the complementary information.

Table 5.2: Definition of GLSZM-based features [2]. Let P be the matrix with size $M \times N$. Scalar $R = \sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j)$. Each element $p(i,j) = P(i,j)/R$.

Feature	Calculation	Description
Small Zone Emphasis	$\sum_{i}^{M}\sum_{j}^{N} \frac{p(i,j)}{j^2}$	Distribution of small zones.
Large Zone Emphasis	$\sum_{i}^{M}\sum_{j}^{N} j^2 p(i,j)$	Distribution of large zones.
Low Gray Level Zone Emphasis	$\sum_{i}^{M}\sum_{j}^{N} \frac{p(i,j)}{i^2}$	Distribution of low gray level values.
High Gray Level Zone Emphasis	$\sum_{i}^{M}\sum_{j}^{N} i^2 p(i,j)$	Distribution of high gray level values.
Small Zone Low Gray Level Emphasis	$\sum_{i}^{M}\sum_{j}^{N} \frac{p(i,j)}{i^2 j^2}$	Joint distribution of small zones and low gray level values.
Small Zone High Gray Level Emphasis	$\sum_{i}^{M}\sum_{j}^{N} \frac{i^2 p(i,j)}{j^2}$	Joint distribution of small zones and high gray level values.
Large Zone High Gray Level Emphasis	$\sum_{i}^{M}\sum_{j}^{N} \frac{j^2 p(i,j)}{i^2}$	Joint distribution of large zones and high gray level values.
Large Zone Low Gray Level Emphasis	$\sum_{i}^{M}\sum_{j}^{N} i^2 j^2 p(i,j)$	Joint distribution of large zones and low gray level values.
Gray Level Non-Uniformity	$\sum_{i}^{M}\left(\sum_{j}^{N} p(i,j)\right)^2$	Similarity of gray level values inside the ROI.
Zone Size Non-Uniformity	$\sum_{j}^{N}\left(\sum_{i}^{M} p(i,j)\right)^2$	Similarity of the size of zones insied the ROI.
Zone Percentage	$R/(jp(i,j))$	homogeneity and distribution of zones inside the ROI.

5.2.3 Improved Evidential Feature Selection

To reduce the complexity of the original EFS discussed in Chapter 3.2.2, a new criterion is constructed for feature selection.

Assuming X_i is a query pattern, other samples in the training pool can be regarded as independent evidence regarding the outcome label of patient i. The evidence offered by

each training instance X_j ($\neq i$) can be quantified as a mass function using (2.8) and (3.4). Since this mass function provides little information when $d_{i,j}$ is too large ($m_{i,j}(\Omega) \approx 1$), it is sufficient to just consider the mass functions offered by the first K (with a large value, e.g., ≥ 10) nearest neighbors of each query pattern X_i.

Let $\{X_{i_1}, \ldots, X_{i_K}\}$ be the selected training samples for X_i. Thus, $\{m_{i,i_1}, \ldots, m_{i,i_K}\}$ are their mass functions. We assign $\{X_{i_1}, \ldots, X_{i_K}\}$ into two different groups (Θ_1 and Θ_2) according to their outcome labels. In each group with the same outcome label, the TBM conjunctive rule (2.4) is used to combine the corresponding mass functions. Hence, when $\Theta_q \neq \emptyset$ ($q = 1$ or 2), the resulting mass function $m_i^{\Theta_q}$ can be represented as

$$\begin{cases} m_i^{\Theta_q}(\{\omega_q\}) & = 1 - \prod\limits_{X_{i_p} \in \Theta_q}^{p=1,\ldots,K} \left(1 - e^{-\gamma_q d_{i,i_p}^2}\right), \\ m_i^{\Theta_q}(\Omega) & = \prod\limits_{X_{i_p} \in \Theta_q}^{p=1,\ldots,K} \left(1 - e^{-\gamma_q d_{i,i_p}^2}\right); \end{cases} \tag{5.1}$$

while, when Θ_q is empty, $m_i^{\Theta_q}(\Omega) = 1$. After that, mass functions $m_i^{\Theta_1}$ and $m_i^{\Theta_2}$ are further combined via the TBM conjunctive rule, so as to obtain a global mass function M_i regarding the class membership of X_i,

$$\begin{cases} M_i(\{\omega_1\}) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\Omega), \\ M_i(\{\omega_2\}) & = m_i^{\Theta_2}(\{\omega_2\}) \cdot m_i^{\Theta_1}(\Omega), \\ M_i(\Omega) & = m_i^{\Theta_1}(\Omega) \cdot m_i^{\Theta_2}(\Omega), \\ M_i(\emptyset) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\{\omega_2\}). \end{cases} \tag{5.2}$$

Based on (3.4), (5.1), and (5.2), M_i, $\forall i \in \{1, \ldots, N\}$, is a function of the binary vector $\Lambda = [\lambda_1, \ldots, \lambda_V]^T$. Quantity $M_i(\emptyset)$ measures the conflict in the neighborhood of X_i. A large $M_i(\emptyset)$ means X_i is locating in a high overlapping area in current feature subspace. Different with $M_i(\emptyset)$, scalar $M_i(\Omega)$ measures the imprecision regarding the class membership of X_i. A large $M_i(\Omega)$ may indicate that X_i is isolated as an outlier from all other training samples in current feature subspace.

According to the requirements of a qualified feature subset described in Chapter 3.2.2, the new loss function with respect to Λ can be defined as

$$L(\Lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{2} \{M_i(\{\omega_q\}) - t_{i,q}\}^2 + \frac{1}{N} \sum_{i=1}^{N} \{M_i(\emptyset)^2 + M_i(\Omega)^2\} + \beta ||\Lambda||_0. \tag{5.3}$$

In (5.3), the first term is a mean squared error measure, where vector t_i is a indicator of the outcome label, with $t_{i,q} = \delta_{i,q}$ if $Y_i = \omega_q$. The second term penalizes feature subsets

that result in high imprecision and large overlaps between different classes. The last term, namely $||\Lambda||_0 = \sum_{v=1}^{V} \lambda_v$, forces the selected feature subset to be sparse. Scalar β (≥ 0) is a hyper-parameter that controls the influence of the sparsity penalty. It should be tuned specifically be a rough gird search strategy.

Considering that the solution of (5.3) is integer constrained (vector Λ should be binary), an integer Genetic Algorithm (GA), namely the MI-LXPM [144], is used to minimize the constructed loss function. As a global optimization algorithm, the MI-LXPM (like other GAs) is more effective than classical optimization methods to find the global optimaL in the case of non-convex problems. The MI-LXPM method mimics biological evolution. At each iteration, it modifies a population of individual feasible solutions according to well-defined selection, crossover and mutation operations, thus producing a new population for the next iteration. Over successive generations (iterations), the population of feasible solutions finally moves toward an optimal solution.

5.2.4 Prior Knowledge

Prior information, such as spatial constraints [176], shape prior [177] and expertise knowledge, is often available in the medical field. In our prediction system, prior knowledge can also be used to guide the feature selection procedure. Since the SUV-based features have shown great significance for assessing the response of a treatment [50,178], we incorporate this important information into EFS as a predefined constraint.

More specifically, a feature ranking method, namely RELIEF [168], is used to rank all kinds of SUV-based features. Let $\tilde{f}$ be a SUV-based feature that exists in each instance X_i, $\forall i \in \{1, \ldots, N\}$. RELIEF assigns a score $S(\tilde{f})$ to $\tilde{f}$ in the form of

$$S(\tilde{f}) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{k} \sum_{j=1}^{k} diff(\tilde{f}, X_i, miss_j^i) - \frac{1}{k} \sum_{j=1}^{k} diff(\tilde{f}, X_i, hit_j^i) \right), \tag{5.4}$$

where hit_j^i and $miss_j^i$, $j \in \{1, \ldots, k\}$, are the nearest neighbors of X_i that originated from the same class and the opposite class, respectively. Function $diff(\tilde{f}, X_1, X_2)$ calculates the difference between the values of the feature $\tilde{f}$ for any two instances X_1 and X_2. The number of nearest neighbors (i.e. k) used in (5.4) is always set to 5 in all our applications.

The obtained score $S(\tilde{f})$ is directly proportional to the informativeness of the feature $\tilde{f}$. Therefore, the SUV-based feature with the largest score is included in EFS as a fixed element of the optimal feature subset. In other words, if the pre-determined feature $\tilde{f}$ is

located in the first dimension of the input feature space, the value of λ_1 is forced to be 1 (can not be 0) when minimizing (5.3). This added constraint drives EFS into a confined searching space. It ensures more robust feature selection, thus increasing the reliability of the prediction system.

5.2.5 Data Balancing

Ensemble with small training sample size, class imbalance is also a typical problem of medical data. Since most of the conventional feature selection methods are designed for well-balanced training data, the class imbalance problem could hinder them to obtain a qualified feature subset. For example, as selecting features according to the accuracy of a specific classifier, SFS and SFFS [139] may output a feature subset that achieves high classification accuracy by simply assigning all training instances to the majority class.

Pre-sampling, either over-sampling the minority class or under-sampling the majority class, is a commonly used approach for the imbalanced learning problems. As a powerful method, Synthetic Minority Over-sampling TEchnique (SMOTE) can generalize the decision region of the minority class via generating synthetic examples [179]. It has shown plenty of successes in many applications, and its variants, such as ADAptive SYNthetic sampling (ADASYN) [56], can further improve the performance.

On this account, ADASYN is adopted in our prediction system to balance the training data for feature selection. The key idea of ADASYN is to adaptively create synthetic samples according to the distribution of the minority class instances, where more instances are generated for the minority class samples that have higher difficulty in learning. The level of difficulty in learning for each minority instance is measured with respect to the ratio of the majority class instances in its k-nearest-neighborhood (k was set to 5 in all our applications). Given an imbalanced training dataset, ADASYN outputs an balanced training dataset via the procedure summarized in Algorithm 2. However, due to the random nature of the data balancing procedure, and also with a limited number of training samples, the balanced training dataset obtained by Algorithm 2 can not always be more representative than the original training dataset. Therefore, in our prediction system, ADASYN is totally executed B (> 1) times to provide B balanced training datasets. EFS is then executed with these balanced datasets to obtain B feature subsets. The final output is determined as the most frequently subset that occurred in the B independent actions.

Algorithm 2: ADASYN-based balancing for feature selection [56].

input : imbalanced dataset $\{(X_i, Y_i)|i = 1, \cdots, N\}$, where $X_i = [x_{i,1}, \cdots, x_{i,V}]^T$ and $Y_i \in \{\omega_1, \omega_2\}$. Assume ω_1 and ω_2 represent the minority class and the majority class, respectively. Let n_{maj} and n_{min} be the number of majority class instances and the number of minority class instances, respectively.

Set the number of synthetic minority class instances as $n_{syn} = n_{maj} - n_{min}$.

for *each sample* X_j *with* $Y_j = \omega_1$ **do**

Find k nearest neighbors of X_j in the training pool.

Calculate the parameter r_j for X_j as $r_j = \Delta_j/k$, where Δ_j is the number of nearest neighbors of X_j that belong to the majority class.

for *each sample* X_j *with* $Y_j = \omega_1$ **do**

Define the level of difficulty in learning for X_j as $\tilde{r}_j = r_j / \sum_{j=1}^{n_{min}} r_j$.

Determine the number of synthetic instances for X_j as $n_j = \tilde{r}_j \times n_{syn}$.

for $l = 1, 2, \ldots, n_j$ **do**

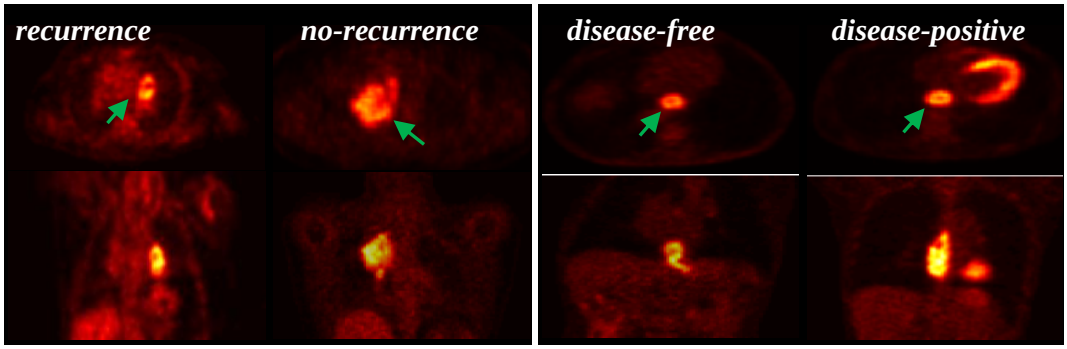
Randomly select a minority class instance, X_r, from the neighbors of X_j.

Randomly generate a scalar $\delta \in [0, 1]$.

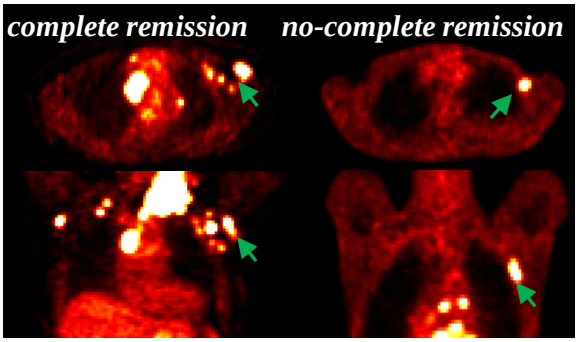
Generate a minority synthetic instance as $S_l^j = X_j + \delta \times (X_r - X_j)$.

5.2.6 Classification

Feature subsets selected by the improved EFS should be used with a classifier to predict the treatment outcome. To this end, case-based methods, such as the K-NN rules and the SVM classifier, are practically good alternatives thanks to their efficiency. As a stable method that offers global treatment of the imperfect knowledge regarding the training data, the EK-NN [87] classification rule, developed in the DST framework, is selected as the default classifier in our prediction system. Parameters used in the EK-NN rule are optimized using the method proposed by [88]. It is worth to note that only the original training dataset with selected features are used to train the classification rule (i.e., no synthetic instance is used during classification), since we assume that instances from the two different classes are widely separated in the feature subspace selected by the improved EFS, while the data balancing procedure has little influence on the classification performance under this circumstance.



(a) Lung Tumors (b) Esophageal Tumors



(c) Lymph Tumors

Figure 5.2: FDG-PET uptakes at tumor staging. For each dataset, two examples with different outcome labels are presented from two complementary views (xy-plane and xz-plane); The arrows point out the tumor locations.

5.3 Clinical Datasets

The prediction system proposed in this paper has been evaluated by three real-world datasets.

1) Lung Tumour Data: A cohort of twenty-five patients with inoperable stage II or III non-small cell lung cancer (NSCLC), treated with curative-intent chemo-radiotherapy (CRT) or radiotherapy (RT). This dataset was extracted from three prospective studies [180]. The total dose of included RT was 60-70 Gy, delivered in daily fractions of 2 Gy and five days a week. Each patient had histological proof of invasive NSCLC, and also had evaluable tumor lesions according to the Response Evaluation Criteria in Solid Tumors (RECIST 1.1). Initial tumor staging was performed based on fibreoptic bronchoscopy, CT scan, pulmonary function tests and biopsy. All patients also underwent FDG-PET scans at initial staging (i.e., PET_0, the baseline). The following PET scans for the same patient were acquired using the same device and under the same operational conditions. The first FDG-PET/CT acquisition (PET_1) was obtained after induction chemotherapy

and before RT, followed by the second FDG-PET/CT scan (PET_2) performed during the fifth week of RT (approximately at a total dose of 40-45 Gy). The treatment response was systematically evaluated and followed-up at three months and one year after RT, or if there was a suspicious relapse. The endpoint was local/distant relapse (LR/DR) vs. complete response (CR) at one year, which was primarily defined by clinical evaluation and CT according to RECIST 1.1, and supplemented by FDG-PET/CT and fiberscope. Finally, nineteen LR/DR patients were grouped into the recurrence class (*majority class*), while the remaining six CR patients were labeled as no-recurrence (*minority class*).

2) Esophageal Cancer Data: A cohort of thirty-six patients with histologically confirmed esophageal squamous cell carcinomas, treated with definitive CRT according to the Herskovic scheme. This dataset was extracted from a retrospective clinical trial [41]. The included RT delivered 2 Gy per fraction per day, five sessions per week for a total of 50 Gy over five weeks. The initial tumor staging was performed based on oesophagoscopy with biopsies, CT scan, and endoscopic ultrasonography. Each patient also underwent a FDG-PET/CT scan at initial tumor staging, but the following PET scans were not complete for all the thirty-six patients. The patients were systematically evaluated and followed-up in a long term up to five years. According to RECIST 1.1 criteria, the response assessment performed one month after CRT was based on clinical evaluation and CT, and possibly supplemented by FDG-PET/CT, and oesophagoscopy with biopsies. Thirteen patients were grouped to the disease-free class (*minority class*), since neither locoregional nor distant disease was detected on them; the remaining twenty-three patients were labeled as disease-positive (*majority class*).

3) Lymph Cancer Data: A cohort of forty-five patients with diffuse large B-cell lymphoma (DLBCL), treated with rituximab and a cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP)/CHOP-like regimen. This dataset was the same as that in [44]. Each patient underwent FDG-PET scans before the onset of chemotherapy (PET_0) and also after three/four cycles of chemotherapy (PET_1). At least three weeks after the end of chemotherapy, the treatment response was evaluated according to the International Workshop Criteria (IWC) for non-Hodgkin lymphoma (NHL) response and according to IWC+PET. Thirty-nine patients were observed complete remission (*majority class*); while, the remaining six patients with refractory or partial response were grouped to the class non-complete remission (*minority class*).

For each dataset, PET image examples acquired at tumor staging are presented in

Figure 5.2.

Feature Description. As discussed in Section 5.2.2, three types of PET image features (SUV-based features, texture features and the temporal changes of them) were extracted. Apart from these image features, variables extracted from other sources of information are also potentially predictive factors. For the esophageal tumor dataset, since only PET images before the treatment were available, some clinical characteristics (patient gender, tumor stage, tumor location, dysphagia grade, etc) were included as the complementary knowledge. In the lymph tumor dataset, only four PET image features were available. As the supplementary information for them, eighteen genes related to the tumor subtype classification, and five genes related to the glucose transportation were also gathered according to the molecular analysis [44]. The three clinical datasets are briefly summarized in Table 5.3, where the number of features and the number of instances are presented. In addition, let the minority (majority) class be the positive (negative) class, we defined the imbalance ratio as $r = N_p/(N_p + N_n)$, where N_p and N_n are the number of positive and negative samples, respectively.

Table 5.3: Description of the three clinical datasets.

dataset	sample size	feature size	imbalance ratio
lung tumor	25	52	0.24
esophageal tumor	36	29	0.36
lymph tumor	45	27	0.13

5.4 Experimental Results

The presented experiments consist of two parts. In the first part, the feature selection performance of the improved EFS was compared with the original EFS, and also compared with some other feature selection methods. In the second part, we assessed the predictive power of the selected feature subsets, and compared them with the predictors that have been proven to be discriminative in clinical studies (e.g., MTV or TLG at staging for the esophageal cancer dataset [41]).

5.4.1 Feature Selection Performance

The improved EFS used in our prediction system was compared with seven other methods, namely two univariate methods (RELIEF and FAST) and five multivariate methods (SFS, SFFS, SVMRFE, KCS, and HFS). As discussed in Section 7.1, the univariate methods rank features according to their individual discriminative power, while the multivariate methods evaluate a subset of features ensemble according to the class separability for a predefined classifier. Because of a limited number of instances, and in order to perform a comprehensive assessment, all the compared methods were evaluated by the Leave-One-Out-Cross-Validation (LOOCV), and also by the .632+ Bootstrapping. At each run of the .632+ Bootstrapping, the learning set is a bootstrap generated by sampling from the studied dataset, while the test set consists of the other samples from the same dataset that do not exit in the bootstrap. Statistically, only 63.2% of the original data are used for learning in each run [181]. The final evaluation is then determined by combining the average performance of all runs (pessimistically biased estimation) with the performance of training and testing both on the original dataset (optimistically biased estimation). The main property of the .632+ Bootstrapping is that it can ensure low biased and variable estimation of classification performance on small-sized datasets [182].

As one of the metrics used to evaluate the selection performance, the robustness of the selected feature subsets was measured by the relative weighted consistency [1]. Its calculation is based on feature occurrence statistics obtained from all iterations of the LOOCV or the .632+ Bootstrapping. The value of the relative weighted consistency ranges between $[0, 1]$, where 1 means all selected feature subsets are approximately identical, while 0 represents no intersection between them. Together with the subset robustness, the classification results obtained during feature selection were also used to assess the feature selection performance. As the most classical figure of merit used in general pattern classification applications, the Accuracy was adopted, which is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5.5}$$

where TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives) represent, respectively, correctly classified positive cases, correctly classified negative cases, incorrectly classified negative cases, and incorrectly classified positive cases. However, only the Accuracy measure is not adequate in the context of clinical management, where the TP rate and the TN rate are more clinically relevant, particularly when instances from

different classes are severely imbalanced. For instance, in cancer diagnosis, there are usually more benign examples (negative cases) than malignant examples (positive cases), while a FN decision (i.e., misclassifying malignant as benign) usually comes at greater costs than a FP decision (i.e., misclassifying benign as malignant). Therefore, to comprehensively assess the classification performance of the imbalanced learning problems, the Receiver Operating Characteristics (ROC) analysis, which was also utilized apart from the Accuracy measure, is more suitable. The ROC makes use of the TP rate and the FP rate, which are defined as

$$TP_{rate} = \frac{TP}{TP + FN}; \quad FP_{rate} = \frac{FP}{TN + FP}. \tag{5.6}$$

In our applications, different pairs of TP_{rate} and FP_{rate} were obtained by applying changing thresholds to the soft classification results, obtained by the EK-NN method, for making a hard decision. Then, based on the ROC curve drawn by all available pairs of TP_{rate} and FP_{rate}, the Area Under the Curve (AUC) was calculated as the complementary measure of the Accuracy in our applications (since all the three examples are imbalanced).

Parameters of all the methods used in sequel are summarized as below:

- For the improved EFS, the parameter B was set to 5. The hyper-parameter β was determined by a rough grid search strategy according to the training performance. The EK-NN classify was adopted to classify training data using feature subsets that obtained by each optional β, and the final value of β was determined as the one that led to the best classification accuracy. On average, good results were obtained with β between $[0.01, 0.07]$ for the lung and lymph tumor datasets, while between $[0.1, 0.3]$ for the esophageal tumor dataset.

- The cutoff thresholds used in RELIEF, FAST and KCS to output selected features were changed from 0.5 to 0.9. Then, the best feature subset was determined according to the average Accuracy. Similarly, the predefined number of selected features that used in SFS, SFFS and SVMRFE was changed from 1 to 5 to output a sparsity feature subset.

- In SFS, SFFS and HFS, the SVM classifier (gaussian kernel, $\sigma = 1$) was chosen as the predefined classifier.

- All parameters used in HFS were the same as that in [22].

- For the compared feature selection methods, the SVM classifier (gaussian kernel, $\sigma = 1$) was adopted to predict the outcome, as it is commonly used with the multivariate methods, and also often used in clinical studies. In our prediction system, the EK-NN classification rule (instead of the SVM classifier) was used with the EFS to predict the treatment outcome.

Table 5.4: Feature selection performance evaluated by the LOOCV. EFS represents our previous work [3], while *i*EFS denotes the improved EFS that proposed in this chapter. "All" represents the results for all the input features (without selection).

	Lung Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	*i*EFS
Robustness	—	0.64	0.65	0.85	0.32	0.56	0.50	**1.00**	0.94	**1.00**
Accuracy	0.76	0.72	0.76	0.88	0.80	0.76	0.84	**1.00**	**1.00**	**1.00**
AUC	0.50	0.60	0.35	0.95	0.61	0.74	0.81	**1.00**	**1.00**	**1.00**
Subset size	52	10	14	2	5	5	3	3	4	4
	Esophageal Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	*i*EFS
Robustness	—	0.94	**1.00**	0.26	0.23	0.80	0.94	0.53	0.92	**1.00**
Accuracy	0.64	0.56	0.64	0.64	0.58	0.72	0.69	0.72	0.83	**0.89**
AUC	0.12	0.54	0.12	0.50	0.55	**0.76**	0.57	0.67	0.69	**0.77**
Subset size	29	2	27	5	5	5	2	5	3	3
	Lymph Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	*i*EFS
Robustness	—	**1.00**	0.85	0.72	0.34	0.64	**1.00**	0.90	0.57	**0.95**
Accuracy	0.87	**0.96**	0.82	0.89	0.87	0.89	**0.96**	0.87	0.89	**0.93**
AUC	0.50	0.68	0.26	0.65	0.29	0.83	0.68	0.36	**0.92**	**0.95**
Subset size	27	1	5	2	5	5	1	4	4	4

5.4.1.1 Evaluation by the LOOCV

The robustness of the selected feature subsets, the average Accuracy, the average AUC, and the average subset size for different methods are summarized in Table 5.4, where the results for all the input features (the SVM classifier was used) are also presented as the baselines

for comparison. From Table 5.4 we can observe that the improved EFS (denoted as ***i***EFS) used in our prediction system always led to robust feature subsets for all the three examples as compared to other methods. Furthermore, it had better (for the esophageal and lymph tumor datasets) or at least the same (for the lung tumor dataset) AUC as compared to other methods. While the Accuracy of the RELIEF and the KCS was slightly better than the proposed ***i***EFS for the lymph tumor dataset (difference of 0.03), the AUC obtained by our method was much better than other methods (minimum difference of 0.12) for this *severely imbalanced example* (imbalanced ratio $r = 0.13$). Comparing the results obtained by the original EFS [3] with the proposed ***i***EFS, it can be found that the data balancing procedure and the incorporated prior knowledge did improve the reliability (relating to robust feature selection) and accuracy (relating to the average Accuracy and AUC) of our prediction system.

5.4.1.2 Evaluation by the .632+ Bootstrapping

The number of Bootstrap samples was set to 100. The robustness of the selected feature subsets, the average Accuracy, the average AUC, and the average subset size are summarized in Table 5.5. Consistent with the results presented in Table 5.4, the robustness of the proposed ***i***EFS that evaluated by the bootstrapping was still better than other methods for all the three examples. In addition, it also led to the best AUC (*especially for the lymph and lung tumor examples with severely imbalanced ratio*) and the best Accuracy. Comparing the results shown in Table 5.5 with that in Table 5.4, we can find that the performance of all the compared methods was declined when evaluated by the bootstrapping. This result is reasonable and foreseeable: Since all the three datasets are small-sized, and due to the random nature of the .632+ bootstrapping, many bootstrap samples may be greatly underrepresented for learning a qualified feature subset. However, it is also worth to note that the difference between the proposed ***i***EFS and other methods was increased under this circumstance, which in some sense confirmed the effectiveness of the proposed method.

5.4.1.3 Selected Feature Subsets

The most frequent feature subsets selected by the improved EFS were kept the same between the LOOCV and the .632+ Bootstrapping for all the three datasets. The detail of the selected features are summarized in Table 5.6 to Table 5.8, respectively. For the lung tumor (Table 5.6), the SUV_{max} during the fith week of RT (PET_2) has also been proven

Table 5.5: Feature selection performance evaluated by the .632+ Bootstrapping. EFS represents our previous work [3], while ***i***EFS denotes the improved EFS that proposed in this paper. "All" represents the results for all the input features (without selection).

	Lung Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	***i***EFS
Robustness	—	0.16	0.11	0.22	0.14	0.12	0.10	0.48	0.21	**0.82**
Accuracy	0.85	0.82	0.82	0.80	0.80	0.84	0.83	0.85	0.81	**0.94**
AUC	0.37	0.64	0.60	0.67	0.66	0.53	0.65	0.81	0.77	**0.94**
Subset size	52	7	10	5	5	5	29	3	4	4
	Esophageal Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	***i***EFS
Robustness	—	0.33	0.61	0.30	0.16	0.31	0.29	0.32	0.44	**0.74**
Accuracy	0.74	0.69	0.74	0.69	0.66	0.74	0.69	0.74	0.77	**0.83**
AUC	0.63	0.66	0.63	0.64	0.63	0.75	0.66	0.71	0.75	**0.82**
Subset size	29	6	25	2	5	5	3	5	3	3
	Lymph Tumor Data									
	All	RELIEF	FAST	SFS	SFFS	SVMRFE	KCS	HFS	EFS	***i***EFS
Robustness	—	**0.56**	0.19	0.25	0.15	0.37	0.33	0.43	0.32	**0.64**
Accuracy	0.92	0.92	0.91	0.90	0.90	0.89	**0.93**	0.91	0.90	**0.93**
AUC	0.62	0.75	0.63	0.73	0.67	0.78	0.77	0.78	0.82	**0.92**
Subset size	27	4	15	1	5	5	2	3	4	4

to have significant predictive power in the clinical study [180]; for the esophageal tumor (Table 5.7), the role of the TLG at tumor staging (PET_0) has been clinically validated in [41]; and for the lymph tumor (Table 5.8), the difference between the SUV_{max} before chemotherapy (PET_0) and the SUV_{max} after three/four cycles of chemotherapy (PET_1) has also been recognized as a variable being capable to predict outcome in [44].

According to above analysis, we could say that the feature subsets determined by our method are in consistent with the predictors that have been verified in clinical studies. More importantly, other kinds of features selected in each subset can give complementary information for these existing measures to improve the prediction performance.

Table 5.6: The most stable feature subset for the lung tumor dataset.

Feature type	Feature description
SUV-based feature	SUV_{max} extracted from PET_2.
GLSZM-based feature	Change of gray-level-non-uniformity between PET_2 and PET_0.
GLSZM-based feature	Change of zone-percentage between PET_1 and PET_0.
GLSZM-based feature	Change of zone-percentage between PET_2 and PET_0.

Table 5.7: The most stable feature subset for the esophageal tumor dataset.

Feature type	Feature description
SUV-based feature	TLG extracted from PET_0.
Clinical characteristic	Tumor staging as II
Clinical characteristic	Patient gender

5.4.2 Prediction Performance

The improved EFS used in our prediction system has robust feature selection performance. To further evaluate the predictive power of these selected feature subsets, the EK-NN classifier with $K = \{1, \ldots, 15\}$ was orderly evaluated by the .632+ Bootstrapping. The number of Bootstrap samples was set to 100. The prediction performance was compared with that obtained by all the input features, and also compared with that obtained by the existing measures (predictors) which have been clinically validated and discussed in the last part of Section 5.4.1. The average AUC with respect to different K is shown in Figure 5.3, where (a)-(c) correspond to the results for the lung tumor, esophageal tumor and lymph tumor dataset, respectively. As can be seen, the selected feature subsets (green line) always led to higher AUC than the input features (blue line) for all the three examples. In addition, they also outperformed the clinically validated predictors (orange line) that self-included in these selected feature subsets. It seems to imply that complementary predictors are well determined for these existing measures in our prediction system.

Misclassified instances: The main reason of misclassification is that the features extracted for these patients are located in the high-overlapping areas in the selected feature space, such as the boundary between two different classes. For the lung tumor dataset, only one

Table 5.8: The most stable feature subset for the lymph tumor dataset.

Feature type	Feature description
SUV-based feature	Change of SUV_{max} between PET_1 and PET_0.
SUV-based feature	SUV_{max} extracted from PET_0.
Gene expression	MME Gene that relates to tumor subtype.
Gene expression	SLC2A5 Gene that relates to glucose transportation.

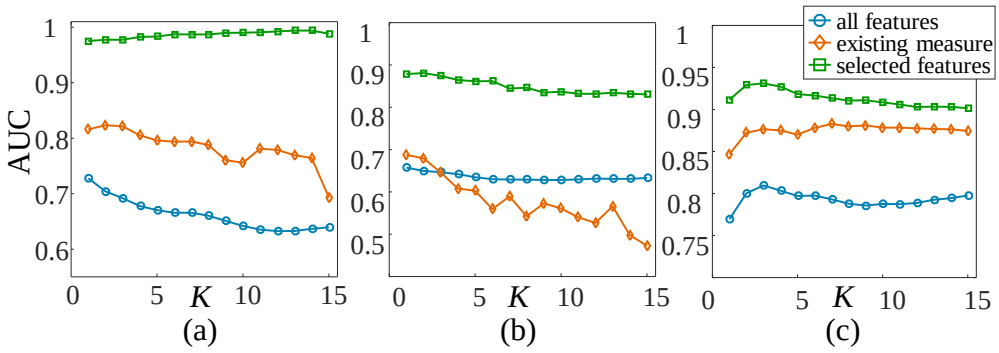


Figure 5.3: Prediction performance of the EK-NN classifier with respect to different K : (a) lung tumor dataset, (b) esophageal tumor dataset, and (c) lymph tumor dataset. "all features", "selected features", and "existing measure" denote the results obtained by the input features, the selected feature subset and the predictor that has been clinically proven, respectively.

patient, which belongs to the recurrence class, was often misclassified; For the lymph tumor dataset, only two instances were frequently misclassified; The prediction performance for the esophageal tumor dataset was poorer than the other two examples, due to the lack of time dependent features extracted from the follow-up PET images.

5.4.3 Discussions

5.4.3.1 Influence of imbalance level

According to the analysis in Section 5.4.1, the competitiveness of the improved EFS seems to be strengthened when the dataset was highly imbalanced (e.g., the lymph tumor example). To support this finding, we further tested our method on a synthetic dataset with respect to different imbalance ratio $r \in \{0.1, 0.2, \ldots, 0.5\}$. Both classes (positive or negative) of this synthetic dataset were generated by multivariate normal distributions. Assume that μ_n and μ_p are the mean vectors for the negative class and the positive class,

respectively; while Σ is the identical covariance matrix for both classes. To be consistent with our clinical examples, the values of μ_n, μ_p and Σ were directly copied as that of the lymph tumor dataset.

Under each level of the imbalance ratio r, 50 samples were generated as a small-sized and imbalanced training dataset. After selecting features using the improved EFS, the EK-NN classifier was learnt to classify a balanced testing dataset. To minimize the uncertainty of the performance estimation, the balanced testing dataset consisted of 3000 test samples, and the evaluation was repeated 50 times for each level of r. The classification results with respect to different imbalance ratio are finally shown in Figure 5.4. As can be seen, Accuracy and AUC obtained by the proposed method are better than directly using all the input features. In particular, the proposed method plays a significant role when the training dataset is severely imbalanced.

5.4.3.2 Role of prior knowledge and data balancing

These two critical modules of our prediction system were successively removed to study the benefits of them. The performance that evaluated by the .632+ Bootstrapping (with 100 Bootstrap Samples) is shown in Figure 5.5, in which ***i***EFS denotes the improved EFS used in our prediction system; while, ***i***EFS$^+$ and ***i***EFS* denote ***i***EFS without data balancing and without prior knowledge, respectively. It can be found that both the included prior knowledge and the data balancing step are helpful for improving the selection performance and the prediction performance. When the dataset is severely imbalance (e.g., the lung tumor example), the data balancing procedure is especially significant for enhancing the robustness and the AUC.

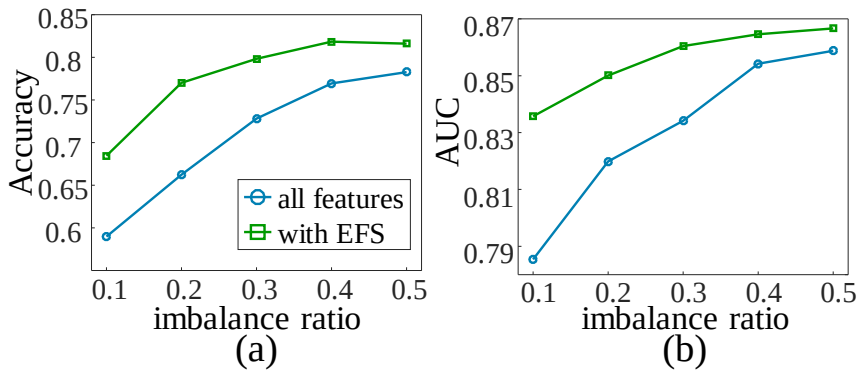


Figure 5.4: (a) Accuracy, and (b) AUC for the synthetic dataset.

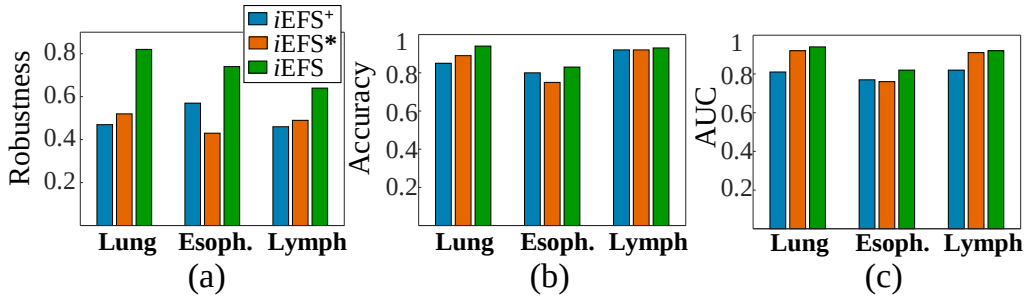


Figure 5.5: (a) Subset robustness, (b) Accuracy, and (c) AUC that evaluated by the .632+ Bootstrapping for the improved EFS without data balancing (iEFS$^+$), the improved EFS without prior knowledge (iEFS*), and the improved EFS (iEFS), respectively.

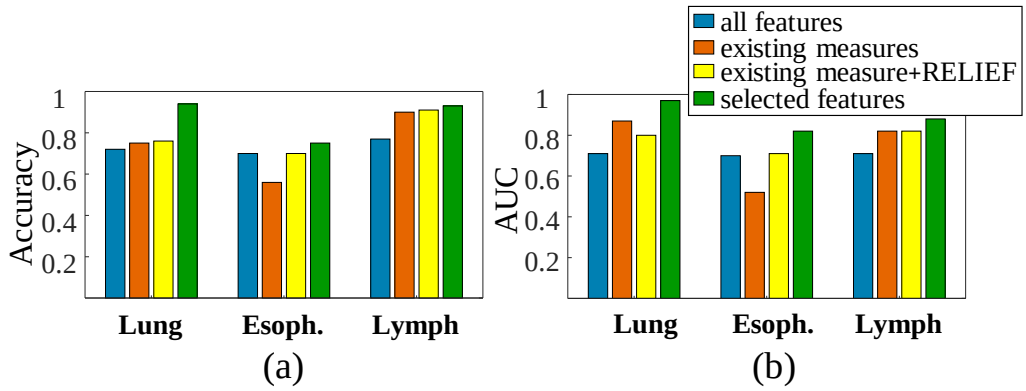


Figure 5.6: (a) Accuracy and (b) AUC of the logistic regression method that evaluated by the .632+ Bootstrapping. The selected features were compared with all the input features, the clinically validated predictors (i.e., existing measures), and the clinically validated predictors joint with features selected by the classical RELIEF (i.e., existing measure+RELIEF).

5.4.3.3 Applicability of the improved EFS

To demonstrate whether the improved EFS has potential benefits for other classifiers (except the EK-NN), the logistic regression, a well-established method widely used in clinical studies, was also adopted to classify the three tumor datasets with the feature subsets detailed in Table 5.6 to Table 5.8. The predictive power of the selected features was compared with that of all the input features, and that of the clinically validated predictors (i.e., existing measures). Additionally, given the clinically validated predictors as the prior, the logistic regression joint with the classical RELIEF, involving to select features to combine with the clinically validated ones, was also presented as the basis for evaluation. Finally, results obtained by the .632+ Bootstrapping (with 100 Bootstrap samples) is summarized in Figure 5.6, based on which we may say that the proposed

method is not only useful for the DST-based classifiers, but also potentially helpful for other classifiers.

5.5 Conclusion

A new framework for PET imaging based cancer treatment outcome prediction has been proposed in this chapter. Features have been extracted from multi-sources of information, which include PET images acquired before and during the treatment, clinical characteristics, and gene expression files. Based on our previous work that has been discussed in Chapter 3, an improved EFS with prior knowledge and data balancing has been proposed to robustly determine the most informative feature subsets from the small-sized and imbalanced training pool. After feature selection, the EK-NN classifier has been trained to predict the outcome. The new prediction system has been evaluated by three clinical studies, showing promising performance with respect to feature selection and classification.

Part III

Automatic Tumor Segmentation in PET Images and PET-CT Images

Accurate delineation of target tumor is indispensable for the practical process of radiation therapy planning. The goal of our study in this part is to develop automatic 3-D algorithms for tumor segmentation in PET and PET-CT images. The uncertainty and imprecision inherent in PET is addressed in the framework of belief functions.

In Chapter 6, an automatic segmentation method based on evidential clustering is proposed. Each image voxel is described not only by intensity but also by complementary voxel-level image features. A spatial regularization based on belief functions is proposed to effectively quantify local homogeneity during the clustering of image voxels. In addition, a specific procedure is adopted in the proposed method to adapt distance measure in unsupervised way, so as to reliably quantifying clustering distortions and neighborhood similarities in the feature space. The method presented in this chapter has been submitted to the IEEE Transactions on Biomedical Engineering.

In Chapter 7, the mono-modality segmentation method proposed in Chapter 6 is further extended to co-segment tumor in PET-CT images. A context term based on belief functions is included in the proposed method to ensure consistent segmentation in PET and CT. During the clustering procedure, the segmentation results in the two distinct mono-modalities are fused via Dempster's combination rule, considering that they contain complementary information for accurate target volume definition.

Spatial-Constrained Evidential Clustering with Adaptive Distance Metric for Tumor Segmentation in PET Images

While the accurate delineation of tumor volumes in FDG-PET is a vital task for diverse objectives in radiation therapy, noise and blur due to the imaging system make it a challenging work. In this chapter, we propose to address the imprecision and noise inherent in PET using Dempster-Shafer theory, a powerful tool for modeling and reasoning with uncertain and/or imprecise information. Based on Dempster-Shafer theory, a novel evidential clustering algorithm is proposed and tailored for the tumor segmentation task in 3D. For accurate clustering of PET voxels, each voxel is described not only by the single intensity value but also complementarily by textural features extracted from a patch surrounding the voxel. Considering that there are a large amount of textures without consensus regarding the most informative ones, and some of the extracted features are even unreliable due to the low-quality PET images, a specific procedure is included in the proposed clustering algorithm to adapt distance metric for properly representing the clustering distortions and the similarities between neighboring voxels. This integrated metric adaptation procedure will realize a low-dimensional transformation from the original space, and will limit the influence of unreliable inputs via feature selection. A Dempster-Shafer-theory-based spatial regularization is also proposed and included in the clustering algorithm, so as to effectively quantify the local homogeneity. The proposed method has been compared with other methods on the real-patient FDG-PET images, showing good performance.

6.1 Introduction

Positron emission tomography (PET), with the radio-tracer fluoro-2-deoxy-D-glucose (FDG), is an advanced imaging tool generally used in clinical oncology for diagnosis, staging, and restaging of tumors. In recent years, FDG-PET has also played an increased important role in adaptive radiation therapy treatment planning process. The goal of adaptive radiation therapy is to improve radiation treatment by incorporating the specificities of individual patient, as well as those of the target tumor, to re-optimize the treatment plan early on during the course of treatment [31]. The utilization of FDG-PET in adaptive radiation therapy has great benefits [15], including 1) as a complement to computed tomography (CT), FDG-PET can help to modify the gross tumor volume (GTV) definition; 2) FDG-PET images can be used to define subvolumes, namely biological target volumes (BTVs), within the tumor target, so as to include tumor biological characteristics in adaptive radiation therapy; 3) some studies, e.g., [19,20,22], have shown that the functional information provided by PET images can predict early the treatment outcome before the end of therapy, offering significant evidence for the adaptation of a more effective treatment plan.

While the accurate delineation of tumor volumes in FDG-PET is a pivotal step for all the purposes discussed above, noisy and blurring images due to the acqusition system make it a challenging work. To this end, diverse automatic or semiautomatic PET image segmentation algorithms have been proposed, which include thresholding methods [61,183], region growing and level set methods [65, 184], statistical methods [67, 185], graph-based methods [28, 70], and clustering methods [75, 80, 186], etc.

As the most commonly used approach owning simple and intuitive nature, thresholding methods usually define a constant [61] or an adaptive [183] threshold value to differentiate lesions from background. The disadvantage is that these methods are sensitive to noise, and have limited performance facing small or heterogeneous positive tissues. Region growing methods also need to select a threshold value as the stopping criterion. To improve the robustness of thresholding segmentation against noise, region growing methods take into account the spatial context in PET images; however, the performance of these methods usually depends heavily on the initialization of the segmentation. Statistical methods assume that positive tissues and surrounding volumes obey different statistical distribution of intensities, e.g., a mixture of Gaussian densities [67]. This kind of method is robust to noise and partial volume effect caused by the low-resolution imaging system; however,

they are sensitive to heterogeneous uptake of positive tissues. Graph-based algorithms, e.g., random walks (RW) [69], can effectively combine global cue with local smoothness by defining foreground and background seeds as hard constraints. Based on previous work with good performance [28,69], an improved version of the classical RW, namely 3D-LARW method [70], has been proposed recently for the segmentation of inhomogeneous or small tumor volumes. The potential disadvantage of these RW methods is that their performance can be influenced by the quality of the seeds.

Clustering methods are suitable for PET image segmentation [57], because the positive tissues are inhomogeneous with varying shapes, which are difficult to be learnt in a supervised manner. In view of the wide application of fuzzy c-means (FCM) clustering in multimodality medical image segmentation tasks [73, 184], Belhassen et al. have proposed a robust approach, called FCM-SW [75], working specifically for the segmentation of heterogeneous tumors in PET images. In the objective function of FCM-SW, the spatial context of image voxels is included for modeling the uncertainty and inaccuracy inherent in PET, thus leading to more stable segmentation than the classical FCM. As an extension of FCM and possibilistic clustering [76], an evidential c-means algorithm (ECM) [78] has been proposed in the framework of Dempster-Shafer theory [77]. A spatial version of ECM, namely SECM [80], has then been proposed recently for lung tumor delineation in multi-tracer PET images. In the objective function of SECM, the local homogeneity is quantified by the weighted sum of the intensity distances from the neighborhood of each voxel to the cluster prototypes. Finding an alternative way to model directly the spatial information in the framework of Dempster-Shafer theory seems to be more appropriate, and may also further enhance the performance of ECM in low-quality PET images.

It is also worth noting that in the clustering methods mentioned above only intensity values have been used to assign voxels into different clusters. Textural features, which describe the spatial environment surrounding each voxel, are very likely to provide complementary information for more accurate segmentation. However, the challenge to include textures in tumor segmentation is that a large amount of textures can be extracted, but there is no consensus regarding the most informative ones; in addition, some of the extracted features may be unreliable or inaccuracy due to the noisy and blurring nature of PET images. Abounding research, e.g., [152,187,188], has shown that learning a distance metric, before or during clustering, adapting to the data at hand could effectively improve the performance of clustering algorithms. However, since the available methods were not designed specifically

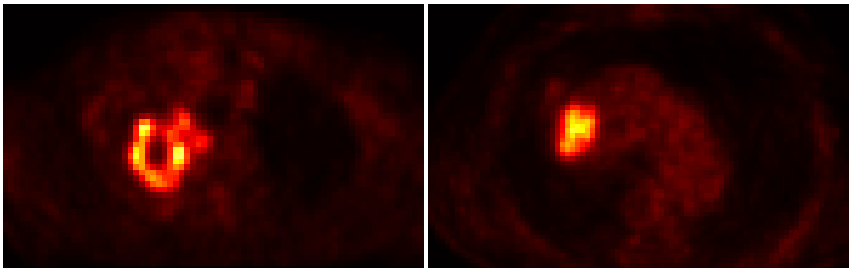


Figure 6.1: Blurring FDG-PET images shown in the axis plane for two different patients, where large intra- and inter-tumor heterogeneity can be observed.

for tackling data that contains unreliable input features, their performance may decline with this kind of imperfect information.

Noise and imprecision modeling is of great concern for reliable PET image segmentation [57], e.g., for the blurring and inhomogeneous positive tissues shown in Figure 6.1. In our study this critical issue is addressed via Dempster-Shafer theory (DST). In the framework of DST, we propose a new ECM clustering algorithm tailored for the delineation of tumor volumes in low-quality 3D PET images. The proposed method has three main objectives: 1) to add textural features as complementary information for the single intensity used in the above methods, so as to obtain more accurate segmentation; 2) to effectively adapt distance metric for well representing the clustering distortions and the similarities between neighboring voxels rather than using directly the simple Euclidean distance. A sparsity constraint is included in the distance metric updating procedure to realize a feature selection via a low-dimensional feature transformation, thus limiting the influence of unreliable input features on the output segmentation; 3) to define a new energy function in the framework of DST using the concept of Markov random field (MRF). By reason that MRF offers a reliable way to consider spatial information [111–114, 189, 190], the new MRF-based energy function is included in the objective function of ECM, and acts as a spatial regularization to effectively quantify the local homogeneity of PET image voxels.

The rest of this chapter is organized as follows. The proposed method is introduced in Section 6.2. In Section 6.3, the proposed method is evaluated by a cohort of real-patient FDG-PET images, and the segmentation performance is compared with that of other methods. Finally, we conclude chapter in Section 6.4.

6.2 Method

Based on the original ECM introduced in Chapter 2.4.2, a new approach, called Evidential c-Means integrating adaptive distance metric and spatial regularization (ECM-MS), is proposed in this section for tumor segmentation in PET.

Let $\{X_1, \ldots, X_n\}$ be a collection of feature vectors in $\mathbb{R}^p$ describing n voxels in a volume of interest (VOI). We assume that all the voxels belong either to the background (i.e. hypothesis ω_1) or to the positive tissue (i.e. hypothesis ω_2), without existence of outliers. Thus, the whole frame of clusters is set as $\Omega = \{\omega_1, \omega_2\}$. Each mass function m satisfies $m(\{\omega_1\}) + m(\{\omega_2\}) + m(\Omega) \equiv 1$, and $m(\emptyset) \equiv 0$. As $m(\Omega)$ measures the ambiguity regarding the clusters ω_1 and ω_2, blurring boundary and severe heterogeneous region will be assigned to $m(\Omega)$.

6.2.1 Spatial Regularization

According to the spatial prior of a PET volume, and as an extension of original 2D MRFs [111–114], the credal partition matrix $\mathbf{M} = \{m_i\}_{i=1}^n$ that we want to learn can be viewed as a specific 3D MRF, where each mass function m_i is a random vector in $\mathbb{R}^3$. Let $\mathbf{\Phi} = \{\Phi(i)\}_{i=1}^n$ be a 3D neighborhood system, where $\Phi(i) = \{1, \ldots, T\}$ is the set of the T neighbors of a voxel i, excluding i. The corresponding masses of voxels in $\Phi(i)$ are $\{m_1^i, \ldots, m_T^i\}$, while the feature vectors of these voxels are $\{X_1^i, \ldots, X_T^i\}$. In the concept of MRF, the distribution of m_i is assumed to be depended on the predefined 3D neighborhood system, i.e., $p(m_i|\{m_j\}_{j\neq i}^n) = p(m_i|\{m_t^i\}_{t\in\Phi(i)})$. Thus, the distribution of $\mathbf{M}$ can be represented as $p(\mathbf{M}) = Z^{-1}\exp\{-U(\mathbf{M})\}$, where Z is a normalizing constant, and $U(\mathbf{M})$ is an *energy function* of the form

$$U(\mathbf{M}) = \eta \sum_{i=1}^{n} \sum_{t\in\Phi(i)} C(i,t), \tag{6.1}$$

where scalar $\eta > 0$ is a tuning parameter, also called the inverse temperature in physics, which controls the degree of local homogeneity in a VOI. The potential function $\sum_{t\in\Phi(i)} C(i,t)$ measures the smoothness around voxel i, in which $C(i,t)$ denotes the inconsistency between voxel i and its neighbor t. In the framework of Dempster-Shafer theory (DST), $C(i,t)$ can be defined as $C(i,t) = \gamma_{it} {d_m}^2(i,t)$, where ${d_m}^2(i,t)$ denotes *the dissimilarity between* m_i *and* m_t^i, while γ_{it} is a weighting factor automatically calculated in feature space.

In this study, the metric defined by Jousselme et al. [191] is adopted to represent the dissimilarity between mass functions of any two adjacent voxels, as it has been commonly used to calculate the conflict between two different pieces of evidence that modeled by DST. As the result, the ${d_m}^2(i,t)$ between m_i and m_t^i, where $t \in \Phi(i)$, is quantified as

$$ {d_m}^2(i,t) = (m_i - m_t^i)\mathbf{Jac}(m_i - m_t^i)^T, \tag{6.2} $$

where **Jac** is a positive definite matrix whose elements are Jaccard indexes, i.e., $\mathbf{Jac}(A,B) = |A \cap B|/|A \cup B|$, $\forall A, B \in 2^\Omega \setminus \emptyset$. The matrix **Jac** used in our study has a specific form, such as

$$ \mathbf{Jac} = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}. \tag{6.3} $$

It is worth noting that (6.2) satisfies the requirements for a valid distance metric. In addition, it effectively accounts for the interaction between the focal elements of Ω [191].

Let m_i, m_1^i, and m_2^i be three masses with the form of

A	$m_i(A)$	$m_1^i(A)$	$m_2^i(A)$
$\{\omega_1\}$	0.8	0.4	0.2
$\{\omega_2\}$	0	0	0.6
Ω	0.2	0.6	0.2

According to this table, $m_i(A)$ is more consistent with $m_1^i(A)$ than with $m_2^i(A)$, as $m_i(A)$ and $m_1^i(A)$ both have mass of belief on $\{\omega_1\}$ and no mass of belief on the opposite hypothesis $\{\omega_2\}$; while m_2^i is strongly concentrated on $\{\omega_2\}$. As a comparison to (6.2), if we quantify the dissimilarities via the simple Euclidean metric, ${d_m}^2(i,1)$ and ${d_m}^2(i,2)$ will inappropriately be identical and equal 0.72. On the contrary, the dissimilarities deduced by (6.2) are $\mathbf{{d_m}^2(i,1) = 0.36}$ and $\mathbf{{d_m}^2(i,2) = 0.72}$, respectively, which measure the distance more reasonably than the Euclidean metric. Therefore, this measure is used to define the specific MRF energy function (6.1). It acts as a spatial regularization to adaptively quantify the local homogeneity during the clustering.

The new objective function of ECM including this MRF-based spatial regularization is proposed as

$$ \mathcal{J}_{ecm}^s(\mathbf{M},\mathbf{V}) = \sum_{i=1}^{n}\sum_{A_j \neq \emptyset} c_j^2 m_{ij}^2 d_{ij}^2 + \eta \sum_{i=1}^{n}\sum_{t \in \Phi(i)} \gamma_{it} {d_m}^2(i,t), \tag{6.4} $$

subject to the constraints $m_{ij} \geq 0$, and

$$\sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} m_{ij} = 1, \quad \forall i = 1, \ldots, n, \tag{6.5}$$

where $\mathbf{V}$ and d_{ij}^2 have the same form as that in (2.10). Matrix $\mathbf{M} = (m_{ij})$ has n rows and 3 columns, in which m_{ij} is the mass of belief attached to the hypothesis that "the object X_i belongs to the credal cluster A_j". The second term of (6.4) is the spatial regularization, in which ${d_m}^2(i,t)$ (i.e. (6.2)) measures the dissimilarity between m_i and m_t^i, while γ_{it} is a weighting factor. The scalar $\eta > 0$ controls the influence of this regularization. It should be predetermined according to the data at hand.

6.2.2 Adaptive Distance Metric

Apart from intensity of voxels, in this study we also attempt to include textural features in ECM as complementary information for more accurate segmentation. The challenge to this is that a large amount of textures can be extracted, but without prior knowledge concerning the most informative features; additionally, these relatively high-dimensional feature vectors are very likely to contain unreliable variables due to the noisy and blurring nature of the PET imaging system. Hence, to obtain a desired segmentation, an adaptive distance metric and feature selection procedure is necessary.

In our previous work presented in Chapter 4, a supervised method has been proposed to learn a low-rank dissimilarity metric for improving the performance of distance-based classifiers on high-dimensional datasets containing unreliable and imprecise features. Distinct from the previous work, and in order to improve the performance of clustering algorithms, here our goal is to adapt distance metric for given data without supervised learning procedure. Therefore, we look for a matrix $\mathbf{D} \in \mathbb{R}^{p \times q}$ during clustering, under the constraint $q \ll p$, by which the dissimilarity between any two feature vectors, say X_1 and X_2, can be represented as

$$d^2(X_1, X_2) = (X_1 - X_2)\mathbf{D}\mathbf{D}^T(X_1 - X_2)^T. \tag{6.6}$$

In other words, matrix $\mathbf{D}$ transforms the original feature space to a low-dimensional subspace, where important input features will have a strong impact when calculating the dissimilarity. To find such a transformation matrix $\mathbf{D}$, the distances d_{ij}^2 used in (6.4) is calculated via (6.6). The spatial regularization that defined by (6.1) is also used to adapt the distance metric. More specifically, for each voxel i and its neighbor t, we define the

weighting factor that used in (6.4) as $\gamma_{it} = (X_i - X_t^i)\mathbf{D}\mathbf{D}^T(X_i - X_t^i)^T$. Then during the minimization of (6.4), a large dissimilarity ${d_m}^2(i,t)$ between m_i and m_t^i will reveal that current distance measure (6.6) is inadequate, and it should be adjusted at the next step to reduce the dissimilarity between X_i and X_t^i, so as to bring the two adjacent voxels closer together.

Based on the above analysis, the objective function (6.4) integrating adaptive distance metric can be updated as

$$\begin{aligned}\mathcal{J}_{ecm}^{ms}(\mathbf{M},\mathbf{V},\mathbf{D}) =& \sum_{i=1}^{n}\sum_{A_j\neq\emptyset} c_j^2 m_{ij}^2 \left[(X_i - \overline{V}_j)\mathbf{D}\mathbf{D}^T(X_i - \overline{V}_j)^T\right] \\ &+\eta\sum_{i=1}^{n}\sum_{t\in\Phi(i)} {d_m}^2(i,t)\left[(X_i - X_t^i)\mathbf{D}\mathbf{D}^T(X_i - X_t^i)^T\right] \\ &+\lambda\left\|\mathbf{D}\right\|_{2,1} - \log\left((\overline{X}_{\omega_1} - \overline{X}_{\omega_2})\mathbf{D}\mathbf{D}^T(\overline{X}_{\omega_1} - \overline{X}_{\omega_2})^T\right),\end{aligned} \tag{6.7}$$

subject to the constraints $m_{ij} \geq 0$ and (6.5). In (6.7), matrix $\mathbf{M} = (m_{ij})$ and $\mathbf{V}$ have the same form as that in (6.4). The dissimilarities between neighboring mass functions, i.e., ${d_m}^2(i,t)$, are still quantified by (6.2). The $\ell_{2,1}$-norm sparsity regularization (i.e. the third term)

$$||\mathbf{D}||_{2,1} = \sum_{i=1}^{p}\sqrt{\sum_{j=1}^{q} D_{i,j}^2} \tag{6.8}$$

is included to select input features during feature transformation. By forcing rows of $\mathbf{D}$ to be zero, this sparsity term only selects the most reliable input features to calculate the linear transformation, thus controlling the influence of unreliable input features on the clustering result. Scalar λ is a hyper-parameter that controls the influence of this regularization. The last term of (6.7) is used to prevent the objective function being trivially solved with $\mathbf{D} = 0$, which collapses all the features vectors into a single point. Vectors $\overline{X}_{\omega_1}$ and $\overline{X}_{\omega_2}$ are two predetermined prototypes (or seeds) for the positive tissue and the background, respectively. A simple and easy initialization of them will be discussed in Section 6.2.3.

Finally, a desired distance metric determined by (6.7) should satisfy 1) neighboring voxels are similar (realizing via the second term), and 2) the tumor seeds and the background seeds are widely separated (realizing via the last term).

6.2.3 Optimization

The objective function defined in (6.7) can be minimized in an EM-like iterative optimization scheme, subject to $m_{ij} \geq 0$, $\forall i \in \{1,\ldots,n\}$ and $j \in \{1,2,3\}$, and to (6.5). Based

on the whole frame of clusters Ω defined at the beginning of this section, the number of clusters c equals 2 in our applications.

6.2.3.1 Initialization

To guide the clustering procedure of ECM-MS, especially to control the integrated metric updating step, we firstly initialize the mass functions and the cluster centers via the original ECM algorithm. A very small number of voxels are then automatically selected as the seeds with predefined cluster labels. More specifically, based on the initial mass functions, image voxels are classified into three credal clusters, i.e., cluster $\{\omega_1\}$, cluster $\{\omega_2\}$, and cluster Ω. To ensure the reliability and to control the number of the selected seeds, the tumor seeds are determined as the voxels whose intensity values are higher than that of the third quartile voxel in the cluster $\{\omega_1\}$; while, the background seeds are determined as the boundary of the VOI. After that, the mass functions for the tumor and background seeds are fixed as $m(\{\omega_1\}) \equiv 1$ and $m(\{\omega_2\}) \equiv 1$, respectively. In addition, the two prototypes, i.e., $\overline{X}_{\omega_1}$ and $\overline{X}_{\omega_2}$, used in (6.7) are calculated as the barycenters of the tumor and background seeds, respectively. The output dimension, namely the number of columns q in $\mathbf{D}$, is then determined by applying principle component analysis on all the feature vectors $\{X_1, \ldots, X_n\}$. The initial $\mathbf{D}$ is constructed by the top 95% eigenvectors.

Then, the optimization procedure alternates between cluster assignment (i.e. $\mathbf{M}$ estimation) in the E-step, and both prototype determination (i.e. $\mathbf{V}$ estimation) and metric adaptation (i.e. $\mathbf{D}$ estimation) in the M-step.

6.2.3.2 E-step

Given $\mathbf{V}$ and $\mathbf{D}$, the minimization of (6.7) only relates to the first two terms, which turns to be a quadratic problem with respect to the mass functions $\mathbf{M} = (m_{ij})$. The derivative of (6.7) concerning the mass function m_i ($\in \mathbb{R}^3$), $\forall i \in \{1, \ldots, n\}$, can be written as

$$\frac{\partial \mathcal{J}_{ecm}^{ms}}{\partial m_i} = 2m_i\mathbf{B} + 2\eta \sum_{j \in \Phi(i)} d_{ij}^2 (m_i - m_j)\mathbf{Jac}, \tag{6.9}$$

where the matrix $\mathbf{Jac}$ is defined by (6.3), $d_{ij}^2 = d^2(X_i, X_j)$ is measured by (6.6), and

$$\mathbf{B} = \begin{pmatrix} c_1^2 d^2(X_i, \overline{V}_1) & 0 & 0 \\ 0 & c_2^2 d^2(X_i, \overline{V}_2) & 0 \\ 0 & 0 & c_3^2 d^2(X_i, \overline{V}_3) \end{pmatrix}, \tag{6.10}$$

Algorithm 3: ECM-MS

Input feature vectors $\{X_1, \ldots, X_n\} \in \mathbb{R}^p$; the spatial neighborhood $\Phi(i)$ of each voxel i; the hyper-parameters η and λ; initial $\mathbf{M}^{(0)}$, $\mathbf{V}^{(0)}$, and $\mathbf{D}^{(0)}$; the tumor and background seeds ;

for $l = 1, 2, \ldots, L$ **do**

E-step: calculate $\mathbf{M}^{(l)}$ using the efficient interior-point algorithm [192] with (6.9), $\mathbf{M}^{(l-1)}$, $\mathbf{V}^{(l-1)}$, and $\mathbf{D}^{(l-1)}$;

M-step I: calculate $\mathbf{V}^{(l)}$ according to (6.11) and $\mathbf{M}^{(l)}$;

M-step II: calculate $\mathbf{D}^{(l)}$ via the Beck-Teboulle proximal gradient algorithm [165] with (6.12), $\mathbf{M}^{(l)}$, $\mathbf{V}^{(l)}$, and $\mathbf{D}^{(l-1)}$;

if *no significant change of* $\mathcal{J}_{ecm}^{ms}$ **then**

break;

Output the final $\mathbf{M}^*$, $\mathbf{V}^*$, and $\mathbf{D}^*$;

where $d^2(X_i, \overline{V}_j)$ is also measured by (6.6). Based on the derivation (6.9), an efficient interior-point algorithm with a limited-memory BFGS approximation of the Hessian matrix [192] is adopted to solve the quadratic problem, so as to obtain the matrix $\mathbf{M}$ at current step.

6.2.3.3 M-step I

The updating of the prototypes is only influenced by the first term of (6.7). Let $f_j = \sum_{i=1}^{n} c_j^2 m_{ij}^2$ and $g_j = \sum_{i=1}^{n} c_j^2 m_{ij}^2 X_i$, $\forall j \in \{1, 2, 3\}$, the centers of the clusters $\{\omega_1\}$ and $\{\omega_2\}$ are calculated, respectively and directly, as

$$\begin{cases} V_1 = \dfrac{2f_2(2g_1 + g_3) + f_3(g_1 - g_2)}{4f_1 f_2 + f_3(f_1 + f_2)}; \\ V_2 = \dfrac{2f_1(2g_2 + g_3) + f_3(g_2 - g_1)}{4f_1 f_2 + f_3(f_1 + f_2)}. \end{cases} \tag{6.11}$$

6.2.3.4 M-step II

It is worth noting that the objective function (6.7), excluding the third term, is differentiable as a function of the transformation matrix $\mathbf{D}$; while, the third term (i.e. the sparsity regularization) is only partly smooth with a singularity at $\mathbf{D} = 0$. For this reason, the proximal Forward-Backward splitting (FBS) algorithms [165] are efficient alternatives to solve the metric updating problem formulated in this step. More precisely, the derivative

of the differentiable part of (6.7) concerning $\mathbf{D}$ can be written as

$$\begin{aligned}\frac{\partial(\cdot)}{\partial \mathbf{D}} &= 2\sum_{i=1}^{n}\sum_{A_j\neq\emptyset} c_j^2 m_{ij}^2 \left[(X_i-\overline{V}_j)^T(X_i-\overline{V}_j)\mathbf{D}\right] \\ &+ 2\eta\sum_{i=1}^{n}\sum_{t\in\Phi(i)} {d_m}^2(i,t)\left[(X_i-X_t^i)^T(X_i-X_t^i)\mathbf{D}\right] \\ &- \frac{2(\overline{X}_{\omega_1}-\overline{X}_{\omega_2})^T(\overline{X}_{\omega_1}-\overline{X}_{\omega_2})\mathbf{D}}{(\overline{X}_{\omega_1}-\overline{X}_{\omega_2})\mathbf{D}\mathbf{D}^T(\overline{X}_{\omega_1}-\overline{X}_{\omega_2})^T},\end{aligned} \tag{6.12}$$

based on which the Beck-Teboulle proximal gradient algorithm [165], an improved version of the classical FBS methods with computational simplify and fast convergence rate, is executed to obtain a required distance metric at current step.

The optimization procedure of the proposed ECM-MS method is briefly summarized in Algorithm 3.

6.2.4 Reducing Uncertainty

Uncertainty is an important issue in PET images due to the inherent noise of the imaging system. During clustering, ECM-MS tackles uncertainty automatically via the integrated spatial regularization and the metric updating procedure. To further reduce the uncertainty after clustering, the mass functions obtained by Algorithm 3 can be post-processed based on DST.

To this end, for each voxel i, the mass functions $\{m_1^i,\ldots,m_T^i\}$ in the 3D neighborhood $\Phi(i)$, i.e., voxels surrounding i, are viewed as T independent pieces of evidence regarding the cluster label of i. We assume that the reliability of each evidence m_t^i is inversely proportional to the spatial distance between i and t. Let this spatial distance be s_{it}^2. Then, based on Dempster's discounting procedure [77], each piece of evidence m_t^i, $\forall t\in\{1,\ldots,T\}$, can be weighted by a coefficient $\mu_t=\exp(-s_{it}^2)$, so as to obtain a discounted mass function

$$\begin{cases} wm_t^i(\{\omega_j\}) &= \mu_t m_t^i(\{\omega_j\}), \quad \forall j=1,2, \\ wm_t^i(\Omega) &= 1-\sum_{j=1}^{2} wm_t^i(\{\omega_j\}). \end{cases} \tag{6.13}$$

Using the Dempster's rule of combination (2.5), the discounted mass functions obtained by (6.13) are fused with the mass function m_i to output a renewed mass function m_i. On the other hand, the above procedure can also be regarded as a filtering operation in a small cubic window.

Finally, a hard partition of the voxels can be obtained via making decision in the plausibility level (2.3) or the pignistic probability level (2.7). Alternatively, a credal partition can be obtained according to the renewed mass functions directly.

6.3 Experiments and Results

In this section, the proposed ECM-MS was evaluated by the FDG-PET images acquired for non-small cell lung cancer (NSCLC) patients. The performance of ECM-MS was compared with that of a constant thresholding method using 40% of the maximum intensity in the lesion (T40%) [61], an adaptive thresholding method (TAD) [183], 3D-LARW [70], FCM-SW [75], SECM [80], and also the original ECM [78].

6.3.1 Material and Features

The FDG-PET images of 14 NSCLC patients were studied. These patients were injected by an average activity of FDG of 261 ± 48 MBq. The obtained PET acquisitions have the same anisotropic resolution of $4.06 \times 4.06 \times 2$ mm^3, and were quantified using standardized uptake values (SUV). The tumor lesions were then manually delineated by experienced clinicians, with the volumes range from 1.9 mL to 135.8 mL. The segmentation was performed in volume of interest (VOI) defined by clinicians.

Considering that the image resolution is anisotropic, a (3×3) window was defined in 2-D to extract features to be used in the proposed ECM-MS. Using this window, the average SUV, the maximum SUV, the minimum SUV, the range of SUV (i.e. maximum$-$minimum), and the standard deviation of SUV were calculated as features for the centering voxel. The gray level size zone matrix (GLSZM) [2] was adopted to extract seven texture features, as its effectiveness in PET image characterization has already been evaluated [20]. Similarly, the gray-level co-occurrence matrix (GLCM) [193] was also utilized to extract fifteen features. To sum up, for each voxel, a 28-dimensional feature vector was extracted, consisting of 6 SUV-based, 7 GLSZM-based, and 15 GLCM-based features.

6.3.2 Evaluation Criteria

Regarding the manually delineation by clinicians as the reference, all the segmentation methods were evaluated by two criteria, i.e., the Dice coefficient (DSC) and the Hausdorff distance (HD). Let S_1 and S_2 be two segmentations with the corresponding boundaries B_1

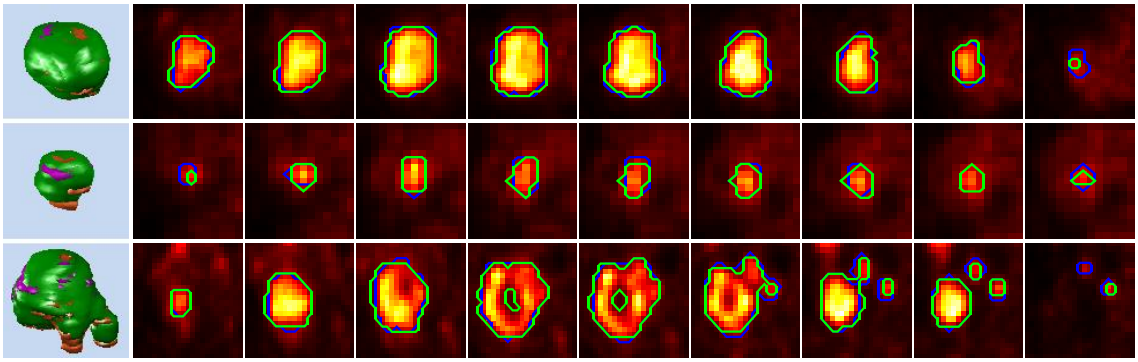


Figure 6.2: Three different tumors delineated by ECM-MS. The first column demonstrates volumes in 3D, where, based on the manually segmentation by clinicians, the green region consists of the true positive and true negative voxels, the magenta region consists of the false positive voxels, while the orange region consists of the false negative voxels. For each tumor volume in the first column, more detailed results, slice by slice in the axial plane, are shown in the following columns correspondingly, where the contours delineated by ECM-MS (green line) are compared with that delineated by clinicians (blue line).

and B_2. Then,

$$DSC = 2|S_1 \cap S_2| / (|S_1| + |S_2|),$$

which measures the overlap between the two different segmentations. While,

$$HD = \max\left\{ \sup_{x \in B_1} \inf_{y \in B_2} d(x,y), \sup_{y \in B_2} \inf_{x \in B_1} d(x,y) \right\},$$

where x and y are points on B_1 and B_2, respectively, and $d(x,y)$ measures the distance (in voxel) between them. Hence, HD quantifies the maximum distance between the boundary points of the two different segmentations.

Table 6.1: The Dice coefficients (DSC) and the Hausdorff Distances (HD) obtained by different segmentation methods on the FDG-PET images for the NSCLC patients. All the results are presented as mean±std.

	T40%	TAD	3D-LARW	ECM	SECM	FCM-SW	ECM-MS
DSC	0.734 ± 0.123	0.716 ± 0.103	0.824 ± 0.07	0.721 ± 0.126	0.767 ± 0.125	0.822 ± 0.11	$\mathbf{0.855 \pm 0.049}$
HD	4.14 ± 4.351	4.224 ± 4.272	4.431 ± 4.516	8.423 ± 3.709	5.886 ± 4.029	4.689 ± 3.733	$\mathbf{2.591 \pm 1.242}$

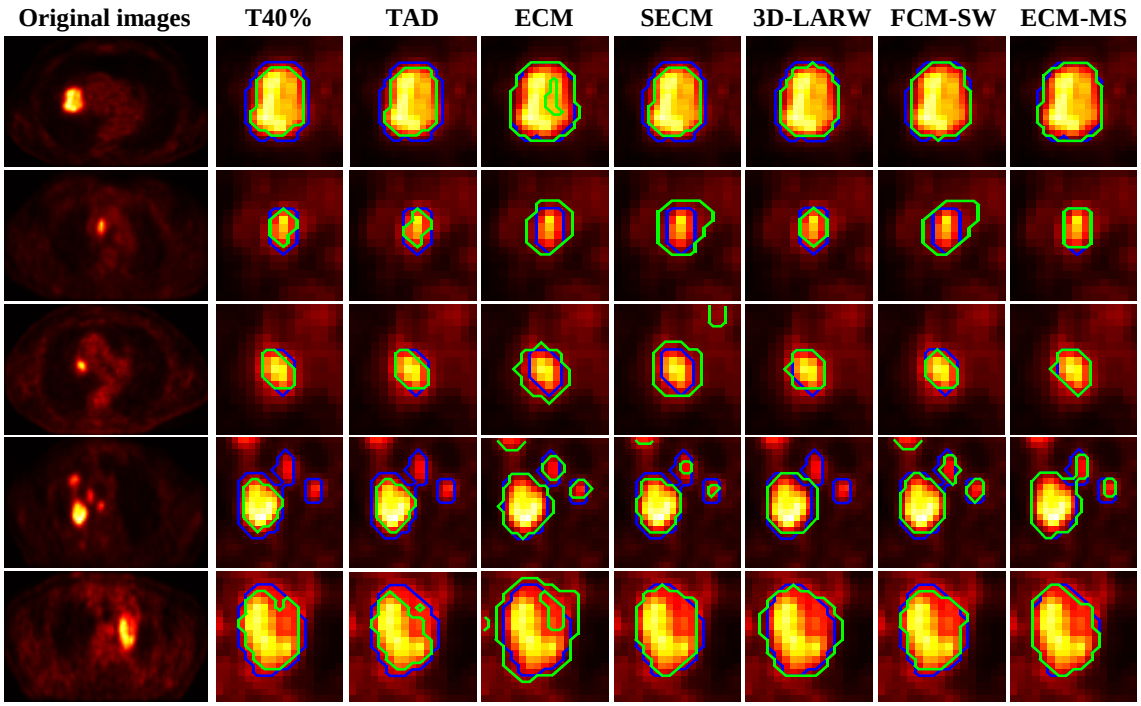


Figure 6.3: Contours delineated by different methods (from the second column to the last column) for five different tumor volumes shown in the axis plane. The first column represents the input images with contours delineated by expert clinicians. The delineation by the seven algorithms (green line) is compared with that by clinicians (blue line) in the following columns.

6.3.3 Results

To demonstrate the performance, as examples, three different PET volumes segmented by ECM-MS are shown in Figure 6.2, where the three rows (from the top to the bottom) correspond to a large tumor, a small tumor, and a heterogenous tumor, respectively. The first column of Figure 6.2 presents the tumor volumes in 3D. Using the manually segmentation by clinicians as the reference, the green region consists of the true positive and true negative voxels, the magenta region consists of the false positive voxels, while the orange region consists of the false negative voxels. For each tumor volume, the second column to the last column of Figure 6.2 show the corresponding results slice by slice in the axis plane (from the top to the bottom), where the green and blue line represent the contours delineated by ECM-MS and clinicians, respectively. As can be seen, the delineation by ECM-MS is in consistent with that by clinicians for all the three examples. It is also worth noting that, for the severely heterogenous tumor shown in the third row, ECM-MS blocked some voxels out from the solid tumor delineated by clinicians. It indicates that

the proposed method may could offer helpful information regarding the radiation necrosis during RT or ART. This property will be discussed in more detail in the next subsection.

The segmentation results obtained by all the methods on the 14 FDG-PET volumes are summarized and compared in Table 6.1, from which we can find that the proposed method obtain better performance, both DSC and HD, than the other six algorithms. To be more comprehensive, the visual examples obtained by these methods are also presented in Figure 6.3 for comparison. The first column of Figure 6.3 presents the axis slices of five different tumors, where the first row is a slice corresponds to a large tumor, the second and the third rows represent two small tumors, while the last two rows represent two heterogenous tumors. The second column to the last column of Figure 6.3 compare the contours delineated by the seven different methods (green line) with that delineated by clinicians (blue line). As can be seen, the contours delineated by the propose method (the last column) are more in consistent with the reference contours in this experiment, especially for the small tumors and heterogenous tumors.

6.3.4 Discussion

6.3.4.1 Uptake Analysis in FDG-PET

In addition to the hard segmentation results presented in Section 6.3.3, ECM-MS can also be adopted to gain deeper insight in the FDG uptake. As an example, a FDG positive tissue, the manual delineation of this tissue by clinicians, and the hard segmentation of this tissue by ECM-MS are shown in Figure 6.4 (a) to (c), respectively. The "credal segmentation" of this tissue by ECM-MS is also presented in Figure 6.4 (d), where the crimson region represents the voxels assigned to the cluster $\{\omega_1\}$ (i.e. the high-uptake voxels); the blue region represents the voxels assigned to the cluster $\{\omega_2\}$ (i.e. the background); while the green region denotes the voxels assigned to the credal cluster Ω, namely the voxels in the blurring boundary of the tumor or the voxels with moderate FDG uptake. Comparing subfigure (c) to (b), we can find that the hard segmentation by ECM-MS blocked some voxels out from the center of the tumor, which seems to be the radiation necrosis. More clear explanation of this result can be obtained from subfigure (d), where the potential radiation necrosis is assigned into the credal cluster Ω as shown in the white circle.

Apart from this credal segmentation result, ECM-MS can also provide detail description of this positive tissue in the mass function level. As shown in Figure 6.5, (a) to (c) represent

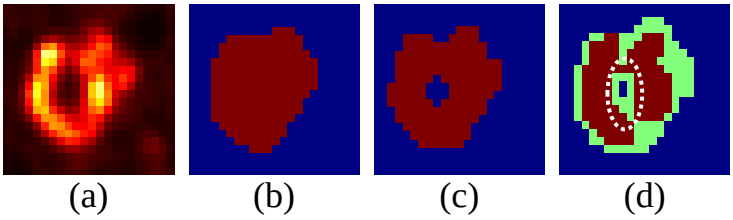


Figure 6.4: (a) A FDG positive tissue; (b) the manual delineation of this tissue by clinicians; (c) the hard segmentation, and (d) the "credal segmentation" of this tissue by ECM-MS. The blue, crimson, and green regions represent, respectively, the segmented background, the segmented high positive tissue, and voxels in the blurring boundary or with moderate FDG uptake.

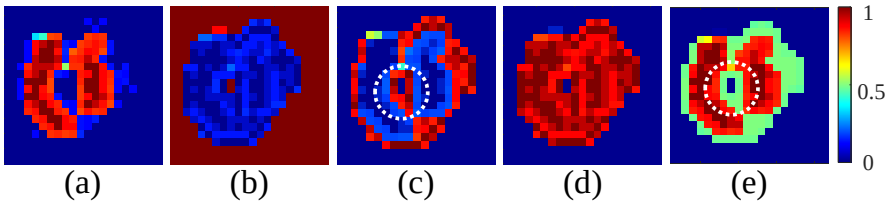


Figure 6.5: (a) to (c) represent the mass functions maps (i.e. $m(\{\omega_1\})$, $m(\{\omega_2\})$, and $m(\{\Omega\})$) obtained by ECM-MS for the positive tissue shown in Figure 6.4 (a); (d) is the plausibility map for the hypothesis of tumor (i.e. $Pl(\{\omega_1\})$); while (e) is the corresponding pignistic probability map (i.e. $BetP(\omega_1)$).

the mass of belief for each voxel attached to the hypothesis of tumor (i.e. $m(\omega_1)$), to the hypothesis of background (i.e. $m(\omega_1)$), and to the whole frame of hypothesis (i.e. $m(\Omega)$), respectively. From subfigure (c) we can find that the blurring boundary and the possible radiation necrosis (in the white circle) of the tumor have higher intensity value (i.e. larger $m(\Omega)$) than other voxels. Using these mass function maps, the plausibility map for the hypothesis of tumor (i.e. $Pl(\{\omega_1\})$ calculated by (2.3)) and the corresponding pignistic probability map defined in [81] can also be complementary deduced as in (d) and (e), respectively. To sum up, based on ECM-MS and Dempster-Shafer theory, we can perform comprehensive analysis of the delineation results.

6.3.4.2 Role of different modules in (6.7)

To evaluate the influence of the spatial regularization, the sparsity regularization, and the uncertainty reduction step (Section 6.2.4) on the final segmentation, we orderly excluded them from ECM-MC. Then, the corresponding results are summarized in Table 6.2, from which we can find that all of them can help to improve the performance, especially the

Table 6.2: Segmentation performance without the spatial regularization (no spatial), the sparse regularization (no $||\mathbf{D}||_{2,1}$), and the uncertainty reduction (no post-processing), respectively.

	No spatial	No $||\mathbf{D}||_{2,1}$	No post-processing	ECM-MS
DSC	0.786 ± 0.089	0.844 ± 0.068	0.848 ± 0.049	$\mathbf{0.855 \pm 0.049}$
HD	3.645 ± 2.526	2.632 ± 1.468	$\mathbf{2.591 \pm 1.242}$	$\mathbf{2.591 \pm 1.242}$

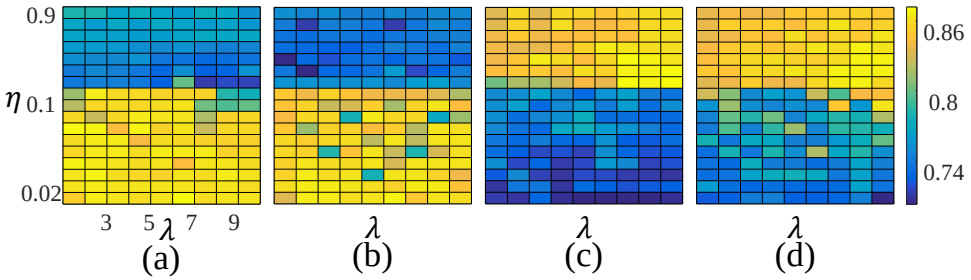


Figure 6.6: The Dice coefficient (i.e. the intensity value) as a function of λ and η. (a) to (c) correspond to four different tumors delineated by clinicians with the volumes of 51.20 mL, 135.80 mL, 18.33 mL and 8.10 mL, respectively.

spatial penalty and the sparsity penalty defined in (6.7).

6.3.4.3 Parameter Setting

The two hyper-parameters utilized in ECM-MS, i.e, η and λ of (6.7), control the influence of the spatial regularization and the influence of the sparsity regularization, respectively. To maximize the segmentation performance, they should be determined taking into account the size of the tumor. As an illustration, we orderly chose a η and a λ from $\{0.01, \ldots, 0.09, 0.1, \ldots, 0.9\}$ and $\{1, \ldots, 10\}$, respectively. Then, ECM-MS was applied on two relatively large tumors (volumes of 51.20 mL and 135.80 mL delineated by clinicians) and two relatively small tumors (volumes of 18.33 mL and 8.10 mL). The segmentation results were finally quantified using DSC, and is summarized in Figure 6.6. It can be found that for the large tumors (i.e. (a) and (b)), ECM-MS had relatively better performance with $\eta \in [0.01, 0.09]$; while, for the two small tumors (i.e. (c) and (d)), $\eta \in [0.1, 0.9]$ is better. Thus, in our experiment, λ was set to 8, while η was set to 0.01 and 0.2, respectively, for large and small tumors.

6.4 Conclusion

In this study, we have investigated to address the imprecision and noise inherent in PET images via Dempster-Shafer theory (DST). Based on DST, an evidential clustering algorithm integrating adaptive distance metric and MRF-based spatial regularization has been proposed for the automatically delineation of tumor volumes in PET images. The experimental results obtained on fourteen stacks of real-patient FDG-PET images have shown the effectiveness of the proposed method. Considering that DST is also widely used for the information fusion task, in the next chapter, we will study how to include the anatomical information provided by CT into the proposed segmentation algorithm, so as to further improve the tumor delineation performance in FDG-PET.

A Robust Evidential Clustering Algorithm Integrating Information Fusion for Co-Segmentation of Tumor in PET-CT Images

The main issue in this chapter is how to fuse complementary information in PET and CT for precise segmentation of tumor. To this end, the mono-modality segmentation method presented in the last chapter is extended to do multi-modality co-segmentation, by concurrently clustering voxels in PET and CT. Under the assumption that tumor contour in PET should be consistent to that in CT, a context term is defined based on belief functions to penalize the difference between segmentations in two mono-modalities. During the iteration of clustering algorithm, the segmentation results in PET and CT are further adjusted by fusing them via Dempster's combination rule, since they are two independent pieces of evidence concerning the definition of target tumor.

The proposed co-segmentation method has been evaluated by fourteen sets of FDG-PET/CT images for Non-Small Cell Lung Cancer (NSCLC) patients, showing good performance.

7.1 Introduction

In the last chapter, an automatic algorithm has been proposed to delineate tumors in PET images in 3-D. Apart from mono-modality segmentation, the development of hybrid PET/CT technique has pointed exciting directions for more accurate tumor volume definition via joint segmentation in co-registered PET-CT images [194]. As compared to PET

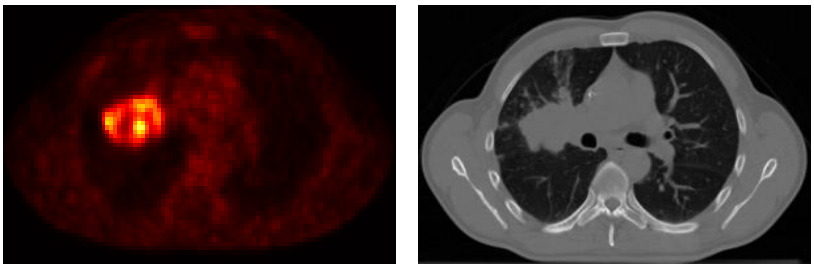


Figure 7.1: A FDG-PET image in the axial plane and the corresponding CT.

(e.g. in Fig. 7.1), while CT images are poor in contrast for differentiating tumor lesions from adjacent normal tissues, they could provide detailed anatomical information that may be complementary to functional information in PET for more comprehensive description of target tumor.

Up to now, most co-segmentation methods in PET and CT are graph-based [26–29]. In [28], based on the construction of a hyper graph, a random walk method was proposed for automatic co-segmentation of multi-modality medical images (e.g. PET-CT and PET-MRI). In [27], which is an extension of [26], the co-segmentation was formulated as a binary labeling problem of Markov Random Field (MRF) on a graph consisting of two sub-graphs. The two sub-graphs correspond to PET and CT images, respectively; the interaction between them was modeled by an adaptive context energy. A maximum flow graph-cut algorithm was then adopted to solve the formulated MRF optimization problem for consistent co-segmentation. In [29], the random walk and graph cut methods were effectively combined, where the former was performed in PET as a stable initialization to improve the co-segmentation performance of the following maximum flow graph cut with a specific energy function. Although the above graph-based methods are efficient and intuitive for the co-segmentation task, the performance of them depends heavily on the quality of predefined tumor and background seeds. In addition, they make decision only according to the information provided by intensity values of image voxels, while other image features (e.g. textures) describing the spatial context of each voxel are very likely to provide complementary information for more reliable tumor delineation.

In this chapter, the previous mono-modality segmentation method based on evidential clustering (Chapter 6) is further extended to perform multi-modality co-segmentation. Tumor volumes in PET and CT are jointly delineated via concurrently clustering image voxels in each mono-modality. A specific context term is defined in the framework of belief functions to penalize the difference between clustering results in PET and CT,

thus driving segmentation in PET to be consistent with that in CT. To effectively taking into complementary information in PET and CT for accurate definition of target tumor, during the minimization of the objective function for evidential clustering, segmentation results in the two mono-modalities are iteratively adjusted by fusing them via Dempster's combination rule (2.5).

The rest of this chapter is organized as follows. The proposed method is introduced in Section 7.2. In Section 7.3, the proposed method is evaluated by a cohort of real-patient PET-CT images, and the segmentation performance is compared with that of other methods. Finally, we conclude chapter in Section 7.4.

7.2 Method

Let $\{X_i^{pt}\}_{i=1}^n$ be feature vectors in $\mathbb{R}^p$ for n voxels in a volume of interest (VOI) of PET, while $\{X_i^{ct}\}_{i=1}^n$ for the corresponding n voxels in CT. We assume that all the voxels, both in PET and CT, belong either to the background (i.e., hypothesis ω_1) or to the positive tissue (i.e., hypothesis ω_2), without existence of outliers. Thus, the whole frame of clusters is set as $\Omega = \{\omega_1, \omega_2\}$. The mass function m for each voxel obeys $m(\{\omega_1\})+m(\{\omega_2\})+m(\Omega) \equiv 1$, and $m(\emptyset) \equiv 0$. As $m(\Omega)$ measures the ambiguity regarding the clusters ω_1 and ω_2, blurring boundary and severe heterogeneous regions will be assigned to $m(\Omega)$.

7.2.1 Cost Function for Joint-Segmentation

The proposed method co-segments tumor in PET and CT via jointly looking for two credal partition matrices $\mathbf{M^{pt}} = \{m_i^{pt}\}_{i=1}^n$ and $\mathbf{M^{ct}} = \{m_i^{ct}\}_{i=1}^n$, where m_i^{pt} and m_i^{ct} ($\in \mathbb{R}^3$) are the mass functions for two corresponding voxels in PET and CT. To this end, a cost consisting of three parts is designed as

$$\mathcal{J}(\mathbf{M^{pt}}, \mathbf{M^{ct}}) = \mathcal{J}_{ecm}^{ms}(\mathbf{M^{pt}}) + \mathcal{J}_{ecm}^{ms}(\mathbf{M^{ct}}) + \gamma \mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}}), \tag{7.1}$$

where $\mathcal{J}_{ecm}^{ms}(\mathbf{M^{pt}})$ and $\mathcal{J}_{ecm}^{ms}(\mathbf{M^{ct}})$ denote, respectively, the independent cost in PET and CT; while, $\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}})$ quantifies the inconsistence between the segmentation in PET and CT; parameter γ controls the influence of this inconsistence. The goal is thus to find a desired pair of $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$ by minimizing the cost (7.1). It is worth indicating that the proposed method encourages consistent co-segmentation in PET and CT.

7.2.1.1 Cost for mono-modality

Both $\mathcal{J}_{ecm}^{ms}(\mathbf{M^{pt}})$ and $\mathcal{J}_{ecm}^{ms}(\mathbf{M^{ct}})$ have the same form as that in (6.7). More specifically, in each mono-modality, the local smoothness is quantified via the MRF-based spatial regularization proposed in Chapter 6.2.1; input image features are selected to adapt distance metric according to the method proposed in Chapter 6.2.2.

7.2.1.2 Cost for inconsistence between PET and CT

The proposed method requires the segmentation in PET and CT to be consistent. Based on this assumption, and using (6.2), the disagreement between the segmentation in PET and CT can then be softly modeled by the dissimilarity between $\mathbf{M^{pt}} = \{m_i^{pt}\}_{i=1}^n$ and $\mathbf{M^{ct}} = \{m_i^{ct}\}_{i=1}^n$. As the result, penalty $\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}})$ in (7.1) is represented by

$$\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}}) = \sum_{i=1}^{n} (m_i^{pt} - m_i^{ct}) \mathbf{Jac} (m_i^{pt} - m_i^{ct})^T, \tag{7.2}$$

where m_i^{pt} and m_i^{ct} are the mass functions of two corresponding voxels in PET and CT, respectively; while, **Jac** is the matrix introduced in (6.3).

7.2.2 Iterative Minimization of the Cost

To find a desired pair of $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$, we propose an iterative scheme to minimize the cost function defined in (7.1), subject to $\sum_{A_j} m_{ij}^{pt} = 1$, and $\sum_{A_j} m_{ij}^{ct} = 1$, $\forall i = 1, \ldots, n$, and $A_j \in \{\{\omega_1\}, \{\omega_2\}, \Omega\}$. The optimization procedure is summarized in Algorithm 4, which can be detailed as follows.

Algorithm 4: Iterative minimization of the cost.

Input feature vectors $\{X_i^{pt}\}_{i=1}^n$ and $\{X_i^{ct}\}_{i=1}^n$; spatial neighborhood $\Phi(i)$ of each voxel i; hyper-parameters η, λ, and γ; initial $\mathbf{M^{pt}}_{(0)}$, $\mathbf{M^{ct}}_{(0)}$, $\mathbf{V^{pt}}_{(0)}$, $\mathbf{V^{ct}}_{(0)}$, $\mathbf{D^{pt}}_{(0)}$, and $\mathbf{D^{ct}}_{(0)}$; tumor and background seeds ;

for $l = 1, 2, \ldots$ **do**

Step 1. Optimization in PET:

- E-step: calculate $\mathbf{M^{pt}}_{(l)}$ using the efficient interior-point algorithm [192] with (7.3), $\mathbf{M^{ct}}_{(l-1)}$, $\mathbf{M^{pt}}_{(l-1)}$, $\mathbf{V^{pt}}_{(l-1)}$, and $\mathbf{D^{pt}}_{(l-1)}$;
- M-step I: calculate $\mathbf{V^{pt}}_{(l)}$ using (6.11) and $\mathbf{M^{pt}}_{(l)}$;
- M-step II: calculate $\mathbf{D^{pt}}_{(l)}$ via the Beck-Teboulle proximal gradient algorithm [165] with (6.12), $\mathbf{M^{pt}}_{(l)}$, $\mathbf{V^{pt}}_{(l)}$, and $\mathbf{D^{pt}}_{(l-1)}$;

Step 2. Optimization in CT:

- E-step: calculate $\mathbf{M^{ct}}_{(l)}$ using the efficient interior-point algorithm [192] with (7.3), $\mathbf{M^{pt}}_{(l)}$, $\mathbf{M^{ct}}_{(l-1)}$, $\mathbf{V^{ct}}_{(l-1)}$, and $\mathbf{D^{ct}}_{(l-1)}$;
- M-step I: calculate $\mathbf{V^{ct}}_{(l)}$ using (6.11) and $\mathbf{M^{ct}}_{(l)}$;
- M-step II: calculate $\mathbf{D^{ct}}_{(l)}$ via the Beck-Teboulle proximal gradient algorithm [165] with (6.12), $\mathbf{M^{ct}}_{(l)}$, $\mathbf{V^{ct}}_{(l)}$, and $\mathbf{D^{ct}}_{(l-1)}$;

Step 3. Update $\mathbf{M^{pt}}_{(l)}$ *via fusing it with* $\mathbf{M^{ct}}_{(l)}$ by (2.5) ;

if *no significant change of (7.1)* **then**

break;

Output qualified $\mathbf{M^{pt}}_*$, and $\mathbf{M^{ct}}_*$;

7.2.2.1 Initialization

We initialize the mass functions (i.e. $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$) and the cluster centers (i.e. $\mathbf{V^{pt}}$ and $\mathbf{V^{ct}}$) for both PET and CT via the original ECM algorithm. Based on the initial mass functions that obtained in PET, a very limited number of tumor and background seeds are determined by the same procedure introduced in Chapter 6.2.3. The locations of tumor and background seeds in PET are then directly copied to CT. After that, the mass functions for the tumor and background seeds are fixed as $m^{pt}(\{\omega_1\}) = m^{ct}(\{\omega_1\}) = 1$, and $m^{pt}(\{\omega_2\}) = m^{ct}(\{\omega_2\}) = 1$, respectively. In addition, the two prototypes, i.e., $\overline{X}^{pt}_{\omega_1}$ ($\overline{X}^{ct}_{\omega_1}$) and $\overline{X}^{pt}_{\omega_2}$ ($\overline{X}^{ct}_{\omega_2}$), used in the last term of (6.7) are calculated as the barycenters of the tumor and background seeds, respectively. Let $\mathbf{D^{pt}}$ and $\mathbf{D^{ct}}$ be low-rank feature transformation matrices to adapt distance metric in PET and CT, respectively. The output dimension, namely the number of columns in $\mathbf{D^{pt}}$ ($\mathbf{D^{ct}}$), is then determined by applying principle component analysis on all the feature vectors $\{X^{pt}_i\}^n_{i=1}$ ($\{X^{ct}_i\}^n_{i=1}$). The initial $\mathbf{D^{pt}}$ ($\mathbf{D^{ct}}$) is constructed by the top 95% eigenvectors.

Then, the optimization procedure alternates between three parts, namely the clustering in PET, the clustering in CT, and the fusion of $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$ at current step. For the clustering in each mono-modality, the optimization iterates between cluster assignment (i.e. $\mathbf{M^{pt}}$ or $\mathbf{M^{ct}}$ estimation) in the E-step, and both prototype determination (i.e. $\mathbf{V^{pt}}$ or $\mathbf{V^{ct}}$ estimation) and distance metric adaptation (i.e. $\mathbf{D^{pt}}$ or $\mathbf{D^{ct}}$ estimation) in the M-step.

7.2.2.2 Optimization in PET

Let $\mathbf{\Phi} = \{\Phi(i)\}^n_{i=1}$ be a 3-D neighborhood system, where $\Phi(i) = \{1, \ldots, T\}$ is the set of the T neighbors of a voxel i in PET, excluding i. The corresponding masses of voxels in $\Phi(i)$ are $\{m^{pt}_{i,1}, \ldots, m^{pt}_{i,T}\}$, while the feature vectors of these voxels are $\{X^{pt}_{i,1}, \ldots, X^{pt}_{i,T}\}$. The optimization in PET only relates to the minimization of the first term and the last term of (7.1), which is performed in an EM-like protocol.

a) E-Step: Given $\mathbf{V^{pt}}$, $\mathbf{D^{pt}}$, and $\mathbf{M^{ct}}$, the minimization of (7.1) turns to be a quadratic problem with respect to the mass functions $\mathbf{M^{pt}} = \{m^{pt}_i\}^n_{i=1}$. The derivative of (7.1) concerning the mass function m^{pt}_i ($\in \mathbb{R}^3$), $\forall i = 1, \ldots, n$, can be written as

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial m^{pt}_i} =& 2m^{pt}_i\mathbf{B} + 2\eta \sum_{t\in\Phi(i)} \left[d^2(X^{pt}_i, X^{pt}_{i,t})\right](m^{pt}_i - m^{pt}_{i,t})\mathbf{Jac} \\ &+ 2\gamma(m^{pt}_i - m^{ct}_i)\mathbf{Jac},\end{aligned} \tag{7.3}$$

where the matrix **Jac** is defined by (6.3), $d^2(X_i^{pt}, X_{i,t}^{pt})$ is measured by (6.6), m_i^{ct} is the mass function for the corresponding ith voxel in CT, and

$$\mathbf{B} = \begin{pmatrix} c_1^2 d^2(X_i^{pt}, \overline{V}_1^{pt}) & 0 & 0 \\ 0 & c_2^2 d^2(X_i^{pt}, \overline{V}_2^{pt}) & 0 \\ 0 & 0 & c_3^2 d^2(X_i^{pt}, \overline{V}_3^{pt}) \end{pmatrix}, \tag{7.4}$$

where $d^2(X_i^{pt}, \overline{V}_j^{pt})$ is also calculated by (6.6). Based on the derivation (7.3), and using $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$ at the last step as initializations, an efficient interior-point algorithm with a limited-memory BFGS approximation of the Hessian matrix [192] is adopted to solve the quadratic problem, so as to obtain the matrix $\mathbf{M^{pt}}$ at current step.

b) M-step I: The updating of the prototypes $\mathbf{V^{pt}}$ is only influenced by the first term of (7.1). More specifically, V_1^{pt} and V_2^{pt} at each iteration are calculated according to (6.11).

c) M-step II: Similar to M-step I, the optimization of $\mathbf{D^{pt}}$ only relates to the first term of (7.1). Based on the gradient information of $\mathbf{D^{pt}}$ that calculated by (6.12), and using $\mathbf{D^{pt}}$ that obtained by the previous iteration as the initialization, the Beck-Teboulle proximal gradient algorithm [165] is adopted to search for a qualified $\mathbf{D^{pt}}$ at current step.

7.2.2.3 Optimization in CT

The adaptation of $\mathbf{M^{ct}}$, $\mathbf{V^{ct}}$, and $\mathbf{D^{ct}}$, which relates to the last two terms of (7.1), is following the same way as that for $\mathbf{M^{pt}}$, $\mathbf{V^{pt}}$, and $\mathbf{D^{pt}}$ discussed above. To update $\mathbf{M^{ct}}$ at current step, $\mathbf{M^{pt}}$ obtained by Section 7.2.2.2 is utilized in (7.3).

7.2.2.4 Fusion of PET and CT

It is worth noting that $\mathbf{M^{pt}}$ and $\mathbf{M^{ct}}$ obtained in Section 7.2.2.2 and 7.2.2.3 are in fact two independent pieces of evidence regarding the same tumor. They are partial complementary, as PET and CT can provide, respectively, functional and anatomical information of the tumor. Therefore, in this step, $\mathbf{M^{pt}}$ is adjusted by the combination of it with $\mathbf{M^{ct}}$ via the Dempster's rule (i.e. (2.5)), which is then used as the initialization for the optimization in PET (i.e. Section 7.2.2.2) of the next iteration. The effectiveness of this step will be further justified in Section 7.3.5.1.

The whole optimization procedure will not terminate the alternation between the steps described in Section 7.2.2.2, 7.2.2.3, and 7.2.2.4, until the value of (7.1) has no significant change between two consecutive iterations.

The mass functions obtained by Algorithm 4 is further processed by the way introduced in Chapter 6.2.4. Finally, a hard partition of the voxels can be obtained via making decision in the plausibility level (2.3), or the pignistic probability level (2.7). Alternatively, a credal partition can be obtained according to the renewed mass functions directly.

7.3 Experiments and Discussions

In this section, the proposed co-segmentation method was evaluated on 14 sets of 3-D FDG-PET/CT images acquired for different non-small cell lung cancer (NSCLC) patients. For quantitative evaluation, the co-segmentation performance was compared with that obtained by each single modality. In addition, the segmentation in PET obtained by the proposed co-segmentation method was also compared with that of other PET segmentation methods, namely 3D-LARW [70], FCM-SW [75], SECM [80], and also the original ECM [78].

7.3.1 Material and Features

The FDG-PET/CT images of 14 NSCLC patients were studied. The PET acquisitions have the same anisotropic resolution of $4.06 \times 4.06 \times 2$ mm^3, and were quantified using standardized uptake values (SUV). The resolution of the corresponding CT images is $0.98 \times 0.98 \times 3$ mm^3. The tumor lesions were manually delineated by experienced clinicians in PET by the guidance of the corresponding CT, with the volumes range from 1.9 mL to 135.8 mL.

Considering that the image resolution is anisotropic, in our experiments, a (3×3) window in 2-D was defined to extract features both in PET and CT images; moreover, for simplicity, the same kinds of features were extracted in both of them. Using the predefined window, the average intensity value, the maximum intensity, the minimum intensity, the range of intensity value (i.e., maximum$-$minimum), and the standard deviation of intensity were calculated as features for the centering voxel. The gray level size zone matrix (GLSZM) [2] was adopted to extract seven texture features, as its effectiveness in medical image characterization has already been evaluated [20]. Similarly, the gray-level co-occurrence matrix (GLCM) [193] was also utilized to extract fifteen features. To sum up, for each PET and CT voxel, a 28-dimensional feature vector was extracted, consisting of 6 intensity-based, 7 GLSZM-based, and 15 GLCM-based features.

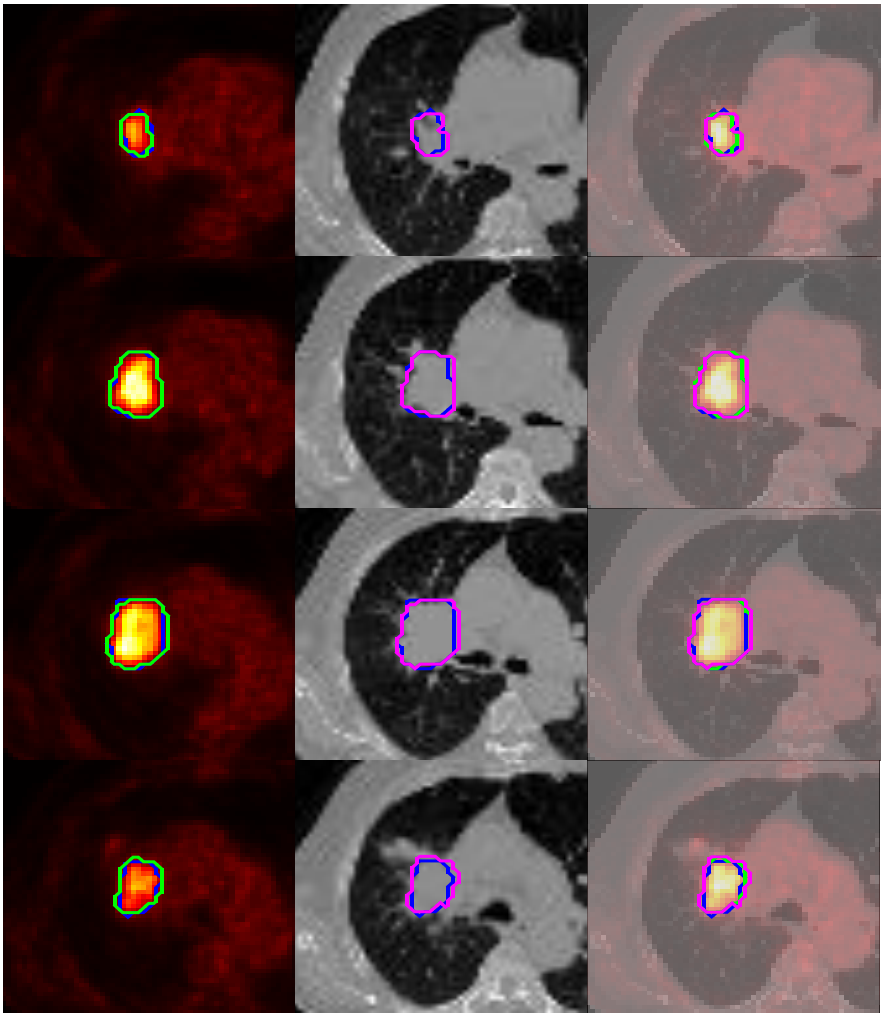


Figure 7.2: A co-segmentation example shown in the axial plane, where contours delineated in PET (green) and CT (magenta) are compared to the ground truth (blue) in the first and the second column, respectively; in the last column, all the contours are overlaid in the fused images.

In consideration of computational costs, after extracting features in PET and CT independently, data in CT were down-sampled to have the same image resolution as PET. Each voxel (and its feature vector) in CT corresponds to one voxel in PET.

7.3.2 Evaluation Criteria

The manual delineation by experienced clinicians was performed on PET images by the guidance of the corresponding CT images. Regarding the manual delineation as the reference, all the segmentation methods were evaluated by two criteria, i.e., the Dice coefficient (DSC) and the Hausdorff distance (HD).

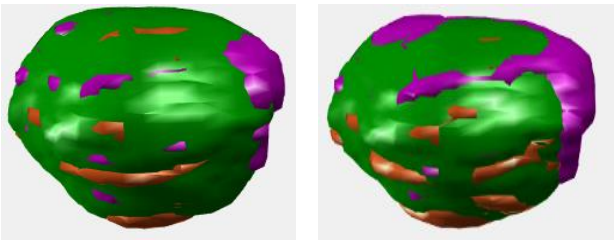


Figure 7.3: Tumor volumes segmented in PET (first column) and CT (second column), where, as compared to the ground truth, where, the green region consists of the true positive and true negative voxels, the magenta region consists of the false positive voxels, while the orange region consists of the false negative voxels.

7.3.3 Parameter Setting

The three hyper-parameters utilized in the proposed method, i.e., η, λ, and γ, control, respectively, the influence of the spatial regularization, the influence of the sparsity regularization, and the consistence of the segmentation in PET and CT. In our experiments, λ and γ were set to 8 and 0.001, respectively; while, to maximize the segmentation performance, the influence of the spatial penalty should be determined by taking into account the size of the tumors under segmentation. More specifically, η was set to 0.003, 0.05, and 0.2 for large, medium and small tumors. The influence of these parameters will be further analyzed in the discussion part (i.e. Section 7.3.5).

7.3.4 Results

7.3.4.1 Illustrative Results of Co-Segmentation

An example to illustrate the co-segmentation performance of the proposed method is shown in Figure 7.2, where each row represents a different slice in the axial plane of the same tumor. In the first and second column, contours delineated in PET (green) and CT (magenta), are compared respectively to the ground truth (blue); while, in the last column, all of them are overlaid in the fused images. As can be seen, segmentation in PET is in consistent with that in CT. The 3D tumor volumes in PET and CT are further shown in Figure 7.3, where the green, magenta, and orange regions represent, respectively, true positive voxels, false positive voxels, and false negative voxels, as compared to the ground truth.

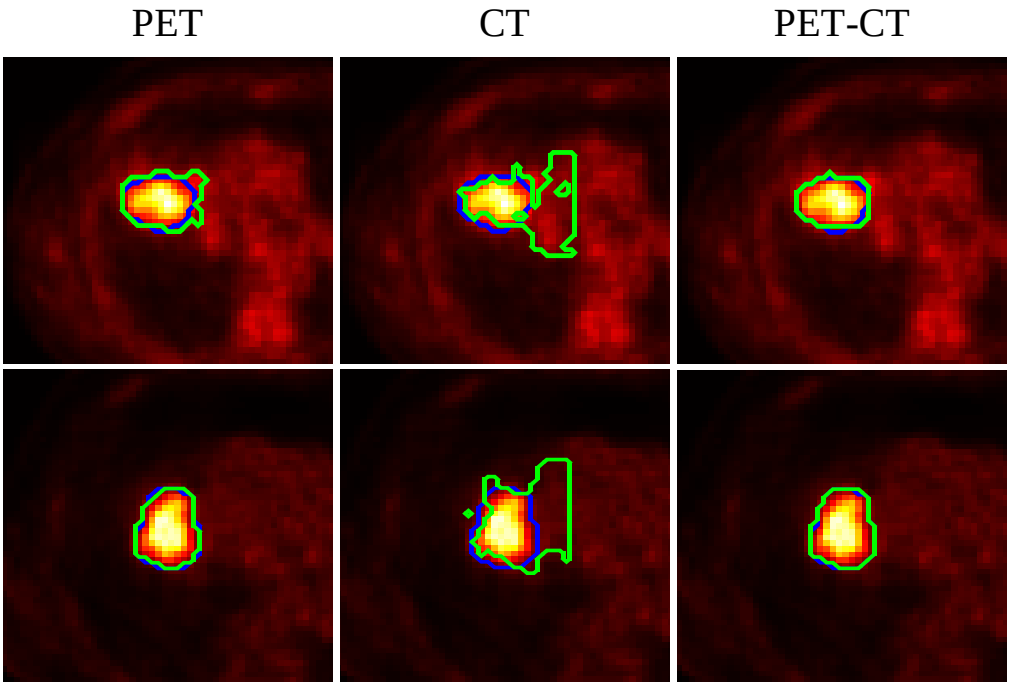


Figure 7.4: Comparing the performance of co-segmentation with the segmentation using single modality. The two rows correspond to two different patients; while the first to the last column represent, respectively, results in PET, results in CT, and results of co-segmentation.

Table 7.1: Average DSC and HD of co-segmentation with that of segmentation using single modality.

	PET only	Co-segment PET	CT only	Co-segment CT
DSC	0.857 ± 0.052	$\mathbf{0.864 \pm 0.043}$	0.260 ± 0.191	$\mathbf{0.855 \pm 0.042}$
HD	2.634 ± 1.236	$\mathbf{2.430 \pm 1.085}$	10.391 ± 1.744	$\mathbf{2.757 \pm 1.631}$

7.3.4.2 Co-Segmentation versus Segmentation in Mono-modality

To demonstrate the effectiveness of the proposed co-segmentation method, its performance was compared with that using mono-modality. In our experiments, segmentation in the two single modalities was performed by setting γ in (7.1) to zero, and removing simultaneously the fusion procedure described in Section 7.2.2.4. Two illustrative results are shown in Figure 7.4, from which we can find that co-segmentation (last column) outperformed mono-modality segmentation in PET (first column) and CT (middle column) in both two cases. The average performance on all the 14 sets of PET/CT images is also summarized in Table 7.1. As can be seen, co-segmentation led to the best DSC and HD in this experiment.

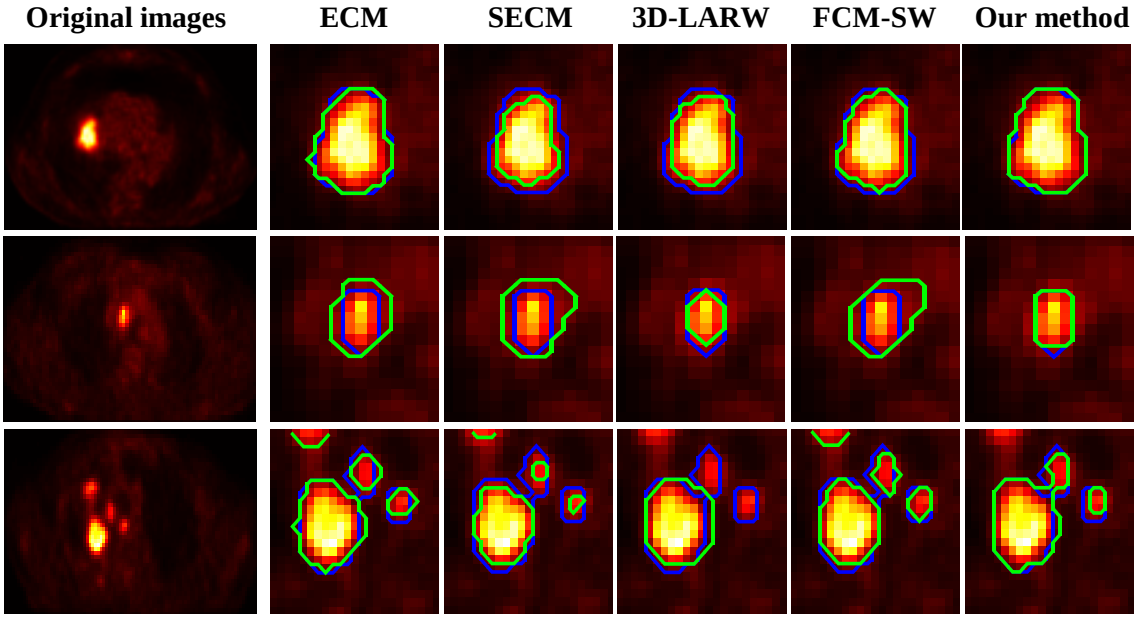


Figure 7.5: Contours delineated by different methods (from the second column to the last column) for three different tumor volumes shown in the axial plane. The first column represents the input images with contours delineated by expert clinicians. The delineation by the five algorithms (green line) is compared with that by clinicians (blue line) in the following columns.

Table 7.2: Quantitative results obtained by different segmentation methods on all the 14 sets of 3D PET/CT images. The DSC and HD are presented as mean±std.

	3D-LARW	**ECM**	**SECM**	**FCM-SW**	**Our method**
DSC	0.824 ± 0.070	0.721 ± 0.126	0.768 ± 0.125	0.822 ± 0.110	$\mathbf{0.864 \pm 0.043}$
HD	4.431 ± 4.516	8.423 ± 3.709	5.886 ± 4.029	4.689 ± 3.734	$\mathbf{2.430 \pm 1.085}$

7.3.4.3 Comparison with Other Methods

The segmentation performance of the proposed method on all the 14 sets of PET/CT images was also compared with that of other four methods, i.e., 3D-LARW [70], the original ECM [78], SECM [80], and FCM-SW [75]. The quantitative comparison is shown in Table 7.2, from which we can find that the proposed method obtained better performance, in terms of both DC and HD, than the other four algorithms. To be more comprehensive, the visual examples obtained by these methods are also presented in Figure 7.5 for comparison. The first column of Figure 7.5 presents the axial slices of three different tumors, where the first row is a slice corresponds to a large tumor, the second row a small tumor, while the last row a heterogenous tumor. The second column to the last column of Figure 7.5 compare

the contours delineated by the five different methods (green line) with that delineated by clinicians (blue line). We can find that the contours delineated by the proposed method (the last column) are more in consistent with the reference contours.

7.3.5 Discussion & Analysis

7.3.5.1 Role of the Fusion via the Dempster's Rule

Based on the Dempster's combination rule (2.5), the fusion of the information from PET and CT (i.e. Section 7.2.2.4), and the fusion of the knowledge from neighboring voxels (i.e. Section 6.2.4) were played a significant role in the proposed method. To evaluate the influence of them, we excluded them from the proposed method. The corresponding segmentation results obtained on all the sets of PET/CT images are then summarized in Table 7.3, from which we can find that the fusion via the Dempster's rule has effectively improved the performance of our method.

Table 7.3: Segmentation performance of the proposed method with/without the fusion procedure based on Dempster's rule.

	without fusion	with fusion
DSC	0.851 ± 0.047	$\mathbf{0.864 \pm 0.043}$
HD	4.424 ± 3.282	$\mathbf{2.430 \pm 1.085}$

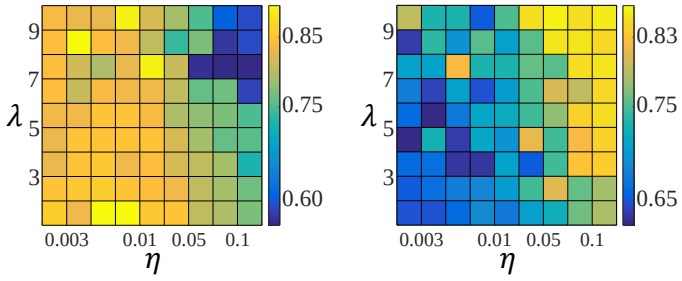


Figure 7.6: The DSC, namely the intensity value, as a function of λ and η. The first and the second column correspond to two tumors with the size of 135.80 mL and 7.60 mL, respectively.

7.3.5.2 Sensitivity to Parameters

As has been introduced in Section 7.3.3, the proposed method is not sensitive to parameter γ and λ; while, to maximize its performance, parameter η of the spatial penalty should be

determined according to the tumor size. As an illustration, we set γ to 0.001, and orderly chose a η and a λ from $\{0.001, \ldots, 0.003, 0.07, 0.01, 0.03, \ldots, 0.07, 0.1, 0.2\}$ and $\{1, \ldots, 10\}$, respectively. Then, the proposed method was applied to segment a relatively large tumor (volume of 135.80 mL) and a relatively small tumor (volume of 7.60 mL). The obtained DSCs are then summarized in Figure 7.6. It can be found that for the large tumor, our method had relative better performance with small η; while, on the contrary, large η is better for the small tumor. Furthermore, we can also find that the value of λ had much less influence than η.

7.4 Conclusion

In this study, the spatial-constrained evidential clustering presented in Chapter 6 has been extended for automatical co-segmentation of tumor in PET-CT images. A specific context term has been included in the proposed method to encourage consistent segmentation between the two distinct mono-modalities. To effectively combine complementary information in PET and CT for accurate definition of target tumor, during the minimization of the constructed cost function, the clustering results in each mono-modality have been iteratively adjusted by fusing them via Dempster's rule. The experimental results have shown that the proposed method performs well as compared to segmentation in each mono-modality. The effectiveness of the included information fusion strategy has also been validated.

Conclusions and Perspectives

Conclusions

Both reliable prediction of treatment outcomes and accurate delineation of tumor volumes are important tasks to ensure the effectiveness of radiation therapy for individual patients. In this thesis, we have investigated how to tackle these two critical challenges:

- We have proposed to predict therapy outcomes primarily using radiomic features extracted from FDG-PET images. In the framework of belief functions, a feature selection method (i.e. EFS in Chapter 3) and a supervised metric learning method (i.e. EDML in Chapter 4) have been proposed to improve the prediction performance of the EK-NN classification rule on uncertain clinical data that contain unreliable input features. To further improve the reliably of our prediction system, the imbalance learning problem on small-sized data, a typical challenge often encountered in the medical domain, has then been carefully dealt with (i.e. Chapter 5).

- We have proposed a robust clustering algorithm based on belief functions for segmenting tumors in PET images (i.e. Chapter 6). Image voxels have been described by intensities and complementary image features for reliably clustering of them. A specific spatial regularization and an unsupervised distance metric updating procedure have been included in the proposed method to effectively quantify, respectively, local homogeneity and clustering distortions. This mono-modality segmentation method has then been extended to co-segment tumors in PET-CT images (i.e. Chapter 7), considering that these two distinct mono-modalities can provide complementary information for accurate definition of target tumor.

Perspectives

We have identified that our work and research presented in this thesis (both cancer therapy outcome prediction and automatic tumor delineation) can be continued and further

improved in multiple directions.

a) The further validation and improvement of the proposed outcome prediction system can be processed from different aspects:

- First of all, the proposed method should be further validated by larger and more clinical datasets.

- In our current work, only FDG-PET based radiomic features have been included to predict treatment outcomes, while other image modalities (e.g. CT and MRI) and PET imaging using other radioactive tracers (e.g. FLT-PET and FMiso-PET) may provide additional information to improve the prediction performance. In order to include multi-sources of information in a unified prediction system, an effective fusion strategy will be a critical and worthy studying issue, considering that these distinct information sources are heterogenous and may contain partial conflicts.

- Considering that deep learning has been applied in diversity fields with great successes, it is interesting and valuable to assess the discriminant power of nonobjective features obtained by deep learning in cancer therapy outcome prediction, and compare their performance with that of traditional image features (e.g. intensities, textures, and shapes, etc.). A foreseeable challenge to apply deep learning in this specific work is that the size of our datasets is too small, which will severely decrease the generalization ability of deep learning. Thus, to tackle this challenge, studying how to process gathered data will be a key issue. Image data augmentation methods can be considered.

- The evidential dissimilarity metric learning method in Chapter 4 has been proposed as a general method without considering the small-sized and imbalanced nature of studied clinical datasets. To further improve its performance in cancer treatment outcome prediction, the imbalanced learning problem should be carefully dealt with. It will be meaningful to assess the performance of other data rebalancing techniques, and compare them with the ADASYN method used in Chapter 5. In addition, it is worth noting that the small-sized nature of clinical datasets may hamper the effectiveness of data rebalancing techniques, since data rebalancing is often randomly, and the quality of synthetic data points depends heavily on the diversity of original training samples. Taking this into account, cost-sensitive learning technique (with-

out data simulation) will be one worth considering solution to further improve the performance of our method on small-sized data.

- The current prediction system presented in Chapter 5 can only tackle binary classification problems. To generalize it for multi-class problems, we can replace (5.2) with

$$\begin{cases} M_i(\{\omega_q\}) & = m_i^{\Theta_q}(\{\omega_q\}) \prod_{p \neq q}^{c} m_i^{\Theta_p}(\Omega), \forall q \in \{1, \ldots, c\} \\ M_i(\Omega) & = \prod_{q=1}^{c} m_i^{\Theta_q}(\Omega) \\ M_i(\emptyset) & = 1 - \sum_{q=1}^{c} M_i(\{\omega_q\}) - M_i(\Omega) \end{cases},$$

and update the first term of the cost function (5.3) as $\frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{c} \{M_i(\{\omega_q\}) - t_{i,q}\}^2$.

b) The proposed mono-modality segmentation and multi-modality co-segmentation methods also needs additional validation and further improvement:

- First of all, the proposed methods should be further validated by more real-patient images, and digital phantom data with objective ground truth and heterogenous uptake. In addition, considering that the proposed methods are unsupervised, it is meaningful to compare their performance with that of segmentation methods based on supervised learning or deep learning.

- The proposed co-segmentation method should be further compared with other published co-segmentation methods.

- In the proposed co-segmentation algorithm, the Dempster's combination rule has been adopted to fuse information from PET and CT to improve the segmentation accuracy. The experimental results presented in Table 7.1 have shown that this strategy can greatly improve the segmentation in CT as compared to mono-modality method, which the improvement in PET is slight. Thus, it is valuable to study other information fusion strategy to further improve the co-segmentation performance.

- In the proposed co-segmentation algorithm, the penalty $\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}})$ defined by (7.2) has been adopted in the cost function (7.1) to encourage consistent segmentation in PET and CT. It quantifies the inconsistence based on the dissimilarity between m_i^{pt} and m_i^{ct}, $\forall i = 1, \ldots, n$, namely mass functions for voxel i in PET and its

correspondent in CT. However, due to different spatial resolution of PET and CT, each voxel in PET may have several potential correspondents in CT, and vice versa. Taking this into account, a more comprehensive $\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}})$ can be defined as

$$\mathcal{J}_{joint}(\mathbf{M^{pt}}, \mathbf{M^{ct}}) = \sum_{i=1}^{n} \sum_{j \in \Phi(i)} (m_i^{pt} - m_j^{ct}) \mathbf{Jac} (m_i^{pt} - m_j^{ct})^T,$$

where $\Phi(i)$ is a set of neighboring voxels with i at its center in CT. In this way, the influence of potential registration errors may be further reduced in the proposed method, which can be further validated in the future.

Bibliography

[1] P. Somol and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1921–1939, 2010.

[2] G. Thibault, J. Angulo, and F. Meyer, "Advanced statistical matrices for texture characterization: application to cell classification," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 630–637, 2014.

[3] C. Lian, S. Ruan, and T. Denœux, "An evidential classifier based on feature selection and two-step classification strategy," *Pattern Recognition*, vol. 48, no. 7, pp. 2318–2327, 2015.

[4] T. Rudroff, J. H. Kindred, and K. K. Kalliokoski, "[18F]-FDG positron emission tomographyąłan established clinical tool opening a new window into exercise physiology," *Journal of Applied Physiology*, vol. 118, no. 10, pp. 1181–1190, 2015.

[5] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[6] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh, "Cancer and radiation therapy: current advances and future directions," *International Journal of Medical Sciences*, vol. 9, no. 3, pp. 193–199, 2012.

[7] E. B. Podgorsak, *Radiation Oncology Physics: A Handbook for Teachers and Students*. IAEA, 2008.

[8] L. A. Dawson and D. A. Jaffray, "Advances in image-guided radiation therapy," *Journal of Clinical Oncology*, vol. 25, no. 8, pp. 938–946, 2007.

[9] L. Xing, B. Thorndyke, E. Schreibmann, Y. Yang, T.-F. Li, G.-Y. Kim, G. Luxton, and A. Koong, "Overview of image-guided radiation therapy," *Medical Dosimetry*, vol. 31, no. 2, pp. 91–112, 2006.

[10] P. Mayles, A. Nahum, and J.-C. Rosenwald, *Handbook of radiotherapy physics: theory and practice*. CRC Press, 2007.

[11] G. C. Pereira, M. Traughber, and R. F. Muzic, "The role of imaging in radiation therapy planning: past, present, and future," *BioMed Research International*, vol. 2014, 2014.

[12] L. Schrevens, N. Lorent, C. Dooms, and J. Vansteenkiste, "The role of PET scan in diagnosis, staging, and management of non-small cell lung cancer," *The Oncologist*, vol. 9, no. 6, pp. 633–643, 2004.

[13] P. Flamen, A. Lerut, E. Van Cutsem, W. De Wever, M. Peeters, S. Stroobants, P. Dupont, G. Bormans, M. Hiele, P. De Leyn, *et al.*, "Utility of positron emission tomography for the staging of patients with potentially operable esophageal carcinoma," *Journal of Clinical Oncology*, vol. 18, no. 18, pp. 3202–3210, 2000.

[14] K. Beal, H. Yeung, and J. Yahalom, "FDG-PET scanning for detection and staging of extranodal marginal zone lymphomas of the MALT type: a report of 42 cases," *Annals of Oncology*, vol. 16, no. 3, pp. 473–480, 2005.

[15] D. Thorwarth, X. Geets, and M. Paiusco, "Physical radiotherapy treatment planning based on functional PET/CT data," *Radiotherapy and Oncology*, vol. 96, no. 3, pp. 317–324, 2010.

[16] Y. Zheng, X. Sun, J. Wang, L. Zhang, X. Di, and Y. Xu, "FDG-PET/CT imaging for tumor staging and definition of tumor volumes in radiation treatment planning in non-small cell lung cancer," *Oncology Letters*, vol. 7, no. 4, pp. 1015–1020, 2014.

[17] S. M. Bentzen and V. Gregoire, "Molecular imaging–based dose painting: A novel paradigm for radiation therapy prescription," in *Seminars in Radiation Oncology*, vol. 21, pp. 101–110, Elsevier, 2011.

[18] S. Ben-Haim and P. Ell, "18F-FDG PET and PET/CT in the evaluation of cancer treatment response," *Journal of Nuclear Medicine*, vol. 50, no. 1, pp. 88–99, 2009.

[19] I. El Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, *et al.*, "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes," *Pattern Recognition*, vol. 42, no. 6, pp. 1162–1171, 2009.

[20] F. Tixier, C. C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.-P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine*, vol. 52, no. 3, pp. 369–378, 2011.

[21] H. Zhang, S. Tan, W. Chen, S. Kligerman, G. Kim, W. D. D'Souza, M. Suntharalingam, and W. Lu, "Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal 18 F-FDG PET features, clinical parameters, and demographics," *International Journal of Radiation Oncology* Biology* Physics*, vol. 88, no. 1, pp. 195–203, 2014.

[22] H. Mi, C. Petitjean, B. Dubray, P. Vera, and S. Ruan, "Robust feature selection to predict tumor treatment outcome," *Artificial Intelligence in Medicine*, vol. 64, no. 3, pp. 195–204, 2015.

[23] M. Vallières, C. Freeman, S. Skamene, I. El Naqa, *et al.*, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine and Biology*, vol. 60, no. 14, p. 5471, 2015.

[24] ICRU, "Prescribing, recording, and reporting photon beam therapy.," *International Commission on Radiation Units and Measurements (ICRU)–Report 50*, 1993.

[25] ICRU, "Prescribing, recording, and reporting photon beam therapy (supplement to ICRU report 50).," *International Commission on Radiation Units and Measurements (ICRU)–Report 62*, 1999.

[26] D. Han, J. Bayouth, Q. Song, A. Taurani, M. Sonka, J. Buatti, and X. Wu, "Globally optimal tumor segmentation in PET-CT images: a graph-based co-segmentation method," in *International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 245–256, Springer, 2011.

[27] Q. Song, J. Bai, D. Han, S. Bhatia, W. Sun, W. Rockey, J. E. Bayouth, J. M. Buatti, and X. Wu, "Optimal co-segmentation of tumor in PET-CT images with context information," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1685–1697, 2013.

[28] U. Bagci, J. K. Udupa, N. Mendhiratta, B. Foster, Z. Xu, J. Yao, X. Chen, and D. J. Mollura, "Joint segmentation of anatomical and functional images: Applications in

quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images," *Medical Image Analysis*, vol. 17, no. 8, pp. 929–945, 2013.

[29] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, "Random walk and graph cut for co-segmentation of lung tumor on PET-CT images," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5854–5867, 2015.

[30] A. Juloori, M. C. Ward, N. P. Joshi, J. F. Greskovich, P. Xia, C. Eric Murray, C. Andrew Dorfmeyer, J. Potter, and S. A. Koyfman, "Adaptive radiation therapy for head and neck cancer," *Applied Radiation Oncology*, vol. 4, no. 3, pp. 12–17, 2015.

[31] D. Yan, F. Vicini, J. Wong, and A. Martinez, "Adaptive radiation therapy," *Physics in Medicine and Biology*, vol. 42, no. 1, p. 123, 1997.

[32] E. Weiss, M. Fatyga, Y. Wu, N. Dogan, S. Balik, W. Sleeman, and G. Hugo, "Dose escalation for locally advanced lung cancer using adaptive radiation therapy with simultaneous integrated volume-adapted boost," *International Journal of Radiation Oncology* Biology* Physics*, vol. 86, no. 3, pp. 414–419, 2013.

[33] X. Geets, M. Tomsej, J. A. Lee, T. Duprez, E. Coche, G. Cosnard, M. Lonneux, and V. Grégoire, "Adaptive biological image-guided IMRT with anatomic and functional imaging in pharyngo-laryngeal tumors: impact on target volume delineation and dose distribution using helical tomotherapy," *Radiotherapy and Oncology*, vol. 85, no. 1, pp. 105–115, 2007.

[34] P. Lambin, R. G. van Stiphout, M. H. Starmans, E. Rios-Velazquez, G. Nalbantov, H. J. Aerts, E. Roelofs, W. van Elmpt, P. C. Boutros, P. Granone, *et al.*, "Predicting outcomes in radiation oncologyąłmultifactorial decision support systems," *Nature Reviews Clinical Oncology*, vol. 10, no. 1, pp. 27–40, 2013.

[35] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, *et al.*, "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[36] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors," *Journal of Nuclear Medicine*, vol. 50, no. Suppl 1, pp. 122S–150S, 2009.

[37] L. Kostakoglu, H. Agress Jr, and S. J. Goldsmith, "Clinical role of FDG PET in evaluation of cancer patients," *Radiographics*, vol. 23, no. 2, pp. 315–340, 2003.

[38] A. van Baardwijk, G. Bosmans, A. Dekker, M. van Kroonenburgh, L. Boersma, S. Wanders, M. Öllers, R. Houben, A. Minken, P. Lambin, *et al.*, "Time trends in the maximal uptake of FDG on PET scan during thoracic radiotherapy. a prospective study in locally advanced non-small cell lung cancer (NSCLC) patients," *Radiotherapy and Oncology*, vol. 82, no. 2, pp. 145–152, 2007.

[39] H. H. Chung, J. W. Kim, K. H. Han, J. S. Eo, K. W. Kang, N.-H. Park, Y.-S. Song, J.-K. Chung, and S.-B. Kang, "Prognostic value of metabolic tumor volume measured by FDG-PET/CT in patients with cervical cancer," *Gynecologic Oncology*, vol. 120, no. 2, pp. 270–274, 2011.

[40] D. E. Soto, M. L. Kessler, M. Piert, and A. Eisbruch, "Correlation between pretreatment FDG-PET biological target volume and anatomical location of failure after radiation therapy for head and neck cancers," *Radiotherapy and Oncology*, vol. 89, no. 1, pp. 13–18, 2008.

[41] C. Lemarignier, F. Di Fiore, C. Marre, S. Hapdey, R. Modzelewski, P. Gouel, P. Michel, B. Dubray, and P. Vera, "Pretreatment metabolic tumour volume is predictive of disease-free survival and overall survival in patients with oesophageal squamous cell carcinoma," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 41, no. 11, pp. 2008–2016, 2014.

[42] P. Vera, S. Mezzani-Saillard, A. Edet-Sanson, J.-F. Ménard, R. Modzelewski, S. Thureau, M.-E. Meyer, K. Jalali, S. Bardet, D. Lerouge, *et al.*, "FDG PET during radiochemotherapy is predictive of outcome at 1 year in non-small-cell lung cancer patients: a prospective multicentre study (RTEP2)," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 41, no. 6, pp. 1057–1065, 2014.

[43] J. van Loon, C. Offermann, M. Öllers, W. van Elmpt, E. Vegt, A. Rahmy, A.-M. C. Dingemans, P. Lambin, and D. De Ruysscher, "Early CT and FDG-metabolic tumour volume changes show a significant correlation with survival in stage I–III small cell lung cancer: a hypothesis generating study," *Radiotherapy and Oncology*, vol. 99, no. 2, pp. 172–175, 2011.

[44] H. Lanic, S. Mareschal, F. Mechken, J.-M. Picquenot, M. Cornic, C. Maingonnat, P. Bertrand, F. Clatot, E. Bohers, A. Stamatoullas, *et al.*, "Interim positron emission

tomography scan associated with international prognostic index and germinal center B cell-like signature as prognostic index in diffuse large B-cell lymphoma," *Leukemia & Lymphoma*, vol. 53, no. 1, pp. 34–42, 2012.

[45] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, *et al.*, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[46] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2015.

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

[48] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014.

[49] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

[50] S. Tan, S. Kligerman, W. Chen, M. Lu, G. Kim, S. Feigenberg, W. D. D'Souza, M. Suntharalingam, and W. Lu, "Spatial-temporal [18 F] FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 85, no. 5, pp. 1375–1382, 2013.

[51] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[52] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, and S. Ruan, "Dealing with uncertainty and imprecision in image segmentation using belief function theory," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 376–387, 2014.

[53] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[54] F. J. Brooks and P. W. Grigsby, "The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake," *Journal of Nuclear Medicine*, vol. 55, no. 1, pp. 37–42, 2014.

[55] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer, 2001.

[56] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

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

[58] J. L. Fox, R. Rengan, W. OąŕMeara, E. Yorke, Y. Erdi, S. Nehmeh, S. A. Leibel, and K. E. Rosenzweig, "Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non–small-cell lung cancer?," *International Journal of Radiation Oncology* Biology* Physics*, vol. 62, no. 1, pp. 70–75, 2005.

[59] J. J. Erasmus, G. W. Gladish, L. Broemeling, B. S. Sabloff, M. T. Truong, R. S. Herbst, and R. F. Munden, "Interobserver and intraobserver variability in measurement of non–small-cell carcinoma lung lesions: Implications for assessment of tumor response," *Journal of Clinical Oncology*, vol. 21, no. 13, pp. 2574–2582, 2003.

[60] H. Zaidi and I. El Naqa, "PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 37, no. 11, pp. 2165–2187, 2010.

[61] Y. E. Erdi, O. Mawlawi, S. M. Larson, M. Imbriaco, H. Yeung, R. Finn, and J. L. Humm, "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," *Cancer*, vol. 80, no. S12, pp. 2505–2509, 1997.

[62] Q. C. Black, I. S. Grills, L. L. Kestin, C.-Y. O. Wong, J. W. Wong, A. A. Martinez, and D. Yan, "Defining a radiotherapy target with positron emission tomography," *International Journal of Radiation Oncology* Biology* Physics*, vol. 60, no. 4, pp. 1272–1282, 2004.

[63] R. Matheoud, P. Della Monica, C. Secco, G. Loi, M. Krengli, E. Inglese, and M. Brambilla, "Influence of different contributions of scatter and attenuation on

the threshold values in contrast-based algorithms for volume segmentation," *Physica Medica*, vol. 27, no. 1, pp. 44–51, 2011.

[64] W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch, "Segmentation of PET volumes by iterative image thresholding," *Journal of Nuclear Medicine*, vol. 48, no. 1, pp. 108–114, 2007.

[65] H. Li, W. L. Thorstad, K. J. Biehl, R. Laforest, Y. Su, K. I. Shoghi, E. D. Donnelly, D. A. Low, and W. Lu, "A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours," *Medical Physics*, vol. 35, no. 8, pp. 3711–3721, 2008.

[66] E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, and M. Miften, "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients," *Medical physics*, vol. 36, no. 10, pp. 4349–4358, 2009.

[67] M. Aristophanous, B. C. Penney, M. K. Martel, and C. A. Pelizzari, "A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography," *Medical Physics*, vol. 34, no. 11, pp. 4223–4235, 2007.

[68] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient ND image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[69] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[70] D. Onoma, S. Ruan, S. Thureau, L. Nkhali, R. Modzelewski, G. Monnehan, P. Vera, and I. Gardin, "Segmentation of heterogeneous or small FDG PET positive tissue based on a 3D-locally adaptive random walk algorithm," *Computerized Medical Imaging and Graphics*, vol. 38, no. 8, pp. 753–763, 2014.

[71] H. Mi, C. Petitjean, P. Vera, and S. Ruan, "Joint tumor growth prediction and tumor segmentation on therapeutic follow-up PET images," *Medical Image Analysis*, vol. 23, no. 1, pp. 84–91, 2015.

[72] D. L. Pham and J. L. Prince, "Adaptive fuzzy segmentation of magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 9, pp. 737–752, 1999.

[73] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data," *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 193–199, 2002.

[74] M. Gong, Y. Liang, J. Shi, W. Ma, and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 573–584, 2013.

[75] S. Belhassen and H. Zaidi, "A novel fuzzy c-means algorithm for unsupervised heterogeneous tumor quantification in PET," *Medical Physics*, vol. 37, no. 3, pp. 1309–1324, 2010.

[76] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[77] G. Shafer, *A mathematical theory of evidence*, vol. 1. Princeton University Press Princeton, 1976.

[78] M.-H. Masson and T. Denœux, "ECM: An evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

[79] N. Makni, N. Betrouni, and O. Colot, "Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: application to prostate multi-parametric MRI," *Information Fusion*, vol. 19, pp. 61–72, 2014.

[80] B. Lelandais, S. Ruan, T. Denœux, P. Vera, and I. Gardin, "Fusion of multi-tracer PET images for dose painting," *Medical Image Analysis*, vol. 18, no. 7, pp. 1247–1259, 2014.

[81] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[82] A. P. Dempster, "Upper and lower probability inferences based on a sample from a finite univariate population," *Biometrika*, vol. 54, no. 3-4, pp. 515–528, 1967.

[83] P. Smets, "The combination of evidence in the transferable belief model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

[84] P. Walley, "Towards a unified theory of imprecise probability," *International Journal of Approximate Reasoning*, vol. 24, no. 2, pp. 125–148, 2000.

[85] H. T. Nguyen, *An introduction to random sets*. CRC press, 2006.

[86] T. Denœux, "40 years of Dempster-Shafer theory," *International Journal of Approximate Reasoning*, vol. 79, pp. 1–6, 2016. 40 years of Research on Dempster-Shafer theory.

[87] T. Denœux, "A K-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[88] L. M. Zouhal and T. Denœux, "An evidence-theoretic K-NN rule with parameter optimization," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 28, no. 2, pp. 263–271, 1998.

[89] T. Denœux, "A neural network classifier based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 30, no. 2, pp. 131–150, 2000.

[90] Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: theoretical foundations," *International Journal of Approximate Reasoning*, vol. 28, no. 2, pp. 91–124, 2001.

[91] J. François, Y. Grandvalet, T. Denoeux, and J.-M. Roger, "Resample and combine: an approach to improving uncertainty representation in evidential pattern classification," *Information Fusion*, vol. 4, no. 2, pp. 75–85, 2003.

[92] Z.-G. Liu, Q. Pan, and J. Dezert, "A new belief-based k-nearest neighbor classification method," *Pattern Recognition*, vol. 46, no. 3, pp. 834–844, 2013.

[93] Z. Liu, Q. Pan, J. Dezert, and G. Mercier, "Credal classification rule for uncertain data based on belief functions," *Pattern Recognition*, vol. 47, no. 7, pp. 2532–2541, 2014.

[94] Z. Liu, Q. Pan, G. Mercier, and J. Dezert, "A new incomplete pattern classification method based on evidential reasoning," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 635–646, 2015.

[95] L. Jiao, Q. Pan, T. Denoeux, Y. Liang, and X. Feng, "Belief rule-based classification system: Extension of FRBCS in belief functions framework," *Information Sciences*, vol. 309, pp. 26–49, 2015.

[96] L. Ma, S. Destercke, and Y. Wang, "Online active learning of decision trees with evidential data," *Pattern Recognition*, vol. 52, pp. 33–45, 2016.

[97] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," *Pattern Recognition*, vol. 52, pp. 85–95, 2016.

[98] T. Denœux and M.-H. Masson, "EVCLUS: evidential clustering of proximity data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 95–109, 2004.

[99] V. Antoine, B. Quost, M.-H. Masson, and T. Denoeux, "CEVCLUS: Evidential clustering with instance-level constraints for relational data," *Soft Computing*, vol. 18, no. 7, pp. 1321–1335, 2014.

[100] Z.-G. Liu, Q. Pan, J. Dezert, and G. Mercier, "Credal c-means clustering method based on belief functions," *Knowledge-Based Systems*, vol. 74, pp. 119–132, 2015.

[101] K. Zhou, A. Martin, Q. Pan, and Z.-G. Liu, "Median evidential c-means algorithm and its application to community detection," *Knowledge-Based Systems*, vol. 74, pp. 69–88, 2015.

[102] K. Zhou, A. Martin, Q. Pan, and Z.-G. Liu, "ECMdd: Evidential c-medoids clustering with multiple prototypes," *Pattern Recognition*, vol. 60, pp. 239–257, 2016.

[103] T. Denœux, "Maximum likelihood estimation from uncertain data in the belief function framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 119–130, 2013.

[104] E. Ramasso and T. Denœux, "Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions.," *IEEE Transactions on Fuzzy Systems*, pp. 1–12, 2013.

[105] Z. Su, Y. Wang, and P. Wang, "Parametric regression analysis of imprecise and uncertain data in the fuzzy belief function framework," *International Journal of Approximate Reasoning*, vol. 54, no. 8, pp. 1217–1242, 2013.

[106] L. Hégarat-Mascle, I. Bloch, D. Vidal-Madjar, *et al.*, "Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 1018–1031, 1997.

[107] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, "Multisensor image segmentation using Dempster-Shafer fusion in Markov fields context," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1789–1798, 2001.

[108] H. Tabia, M. Daoudi, J.-P. Vandeborre, and O. Colot, "A new 3D-matching method of nonrigid and partially similar models using curve analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 852–858, 2011.

[109] Z. Liu, J. Dezert, G. Mercier, and Q. Pan, "Dynamic evidential reasoning for change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1955–1967, 2012.

[110] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 406–417, 2014.

[111] W. Pieczynski and D. Benboudjema, "Multisensor triplet Markov fields and theory of evidence," *Image and Vision Computing*, vol. 24, no. 1, pp. 61–69, 2006.

[112] M. E. Y. Boudaren, L. An, and W. Pieczynski, "Unsupervised segmentation of sar images using gaussian mixture-hidden evidential markov fields," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1865–1869, 2016.

[113] M. E. Y. Boudaren, L. An, and W. Pieczynski, "Dempster-Shafer fusion of evidential pairwise Markov fields," *International Journal of Approximate Reasoning*, vol. 74, pp. 13–29, 2016.

[114] M. E. Y. Boudaren and W. Pieczynski, "Dempster-Shafer fusion of evidential pairwise Markov chains," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 6, pp. 1598–1610, 2016.

[115] Y. Bi, J. Guan, and D. Bell, "The combination of multiple classifiers using an evidential reasoning approach," *Artificial Intelligence*, vol. 172, no. 15, pp. 1731–1751, 2008.

[116] T. Denœux, "Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence," *Artificial Intelligence*, vol. 172, no. 2, pp. 234–264, 2008.

[117] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denœux, "Multimodal information fusion for urban scene understanding," *Machine Vision and Applications*, pp. 1–19, 2014.

[118] R. R. Yager, "On the Dempster-Shafer framework and new combination rules," *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[119] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," *Computational Intelligence*, vol. 4, no. 3, pp. 244–264, 1988.

[120] C. K. Murphy, "Combining belief functions when evidence conflicts," *Decision Support Systems*, vol. 29, no. 1, pp. 1–9, 2000.

[121] Z. Liu, J. Dezert, Q. Pan, and G. Mercier, "Combination of sources of evidence with different discounting factors based on a new dissimilarity measure," *Decision Support Systems*, vol. 52, no. 1, pp. 133–141, 2011.

[122] T. Denoeux, "Analysis of evidence-theoretic decision rules for pattern classification," *Pattern Recognition*, vol. 30, no. 7, pp. 1095–1107, 1997.

[123] B. Quost, M.-H. Masson, and T. Denœux, "Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 353–374, 2011.

[124] H. Altınçay, "Ensembling evidential k-nearest neighbor classifiers through multi-modal perturbation," *Applied Soft Computing*, vol. 7, no. 3, pp. 1072–1083, 2007.

[125] M.-H. Masson and T. Denœux, "Ensemble clustering in the belief functions framework," *International Journal of Approximate Reasoning*, vol. 52, no. 1, pp. 92–109, 2011.

[126] E. Côme, L. Oukhellou, T. Denœux, and P. Aknin, "Learning from partially supervised data using mixture models and belief functions," *Pattern recognition*, vol. 42, no. 3, pp. 334–348, 2009.

[127] M.-H. Masson and T. Denœux, "RECM: Relational evidential c-means algorithm," *Pattern Recognition Letters*, vol. 30, no. 11, pp. 1015–1026, 2009.

[128] T. Denœux and P. Smets, "Classification using belief functions: relationship between case-based and model-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 6, pp. 1395–1406, 2006.

[129] E. Fix and J. L. Hodges Jr, "Discriminatory analysis-nonparametric discrimination: consistency properties," tech. rep., DTIC Document, 1951.

[130] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural computation*, vol. 1, no. 4, pp. 425–464, 1989.

[131] L. Olshen and C. J. Stone, "Classification and regression trees," *Wadsworth International Group*, 1984.

[132] Z. Liu, Q. Pan, and J. Dezert, "Evidential classifier for imprecise data based on belief functions," *Knowledge-Based Systems*, vol. 52, pp. 246–257, 2013.

[133] J. Yang and D. Xu, "Evidential reasoning rule for evidence combination," *Artificial Intelligence*, vol. 205, pp. 1–29, 2013.

[134] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[135] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[136] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[137] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.

[138] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[139] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[140] S. Nakariyakul and D. P. Casasent, "An improvement on floating search algorithms for feature subset selection," *Pattern Recognition*, vol. 42, no. 9, pp. 1932–1940, 2009.

[141] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[142] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *The Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[143] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer methods in applied mechanics and engineering*, vol. 186, no. 2, pp. 311–338, 2000.

[144] K. Deep, K. P. Singh, M. Kansal, and C. Mohan, "A real coded genetic algorithm for solving integer and mixed integer optimization problems," *Applied Mathematics and Computation*, vol. 212, no. 2, pp. 505–518, 2009.

[145] S. Perkins, K. Lacker, and J. Theiler, "Grafting: Fast, incremental feature selection by gradient descent in function space," *The Journal of Machine Learning Research*, vol. 3, pp. 1333–1356, 2003.

[146] A. Frank and A. Asuncion, "Uci machine learning repository, 2010," *URL http://archive. ics. uci. edu/ml*, vol. 15, p. 22, 2011.

[147] C. M. Bishop, *Neural networks for pattern recognition*. Oxford University Press, 1995.

[148] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[149] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 325–327, 1976.

[150] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, 2006.

[151] D. G. Lowe, "Similarity metric learning for a variable-kernel classifier," *Neural Computation*, vol. 7, no. 1, pp. 72–85, 1995.

[152] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, pp. 505–512, 2002.

[153] D.-Y. Yeung and H. Chang, "A kernel approach for semisupervised metric learning," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 141–149, 2007.

[154] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216, ACM, 2007.

[155] W. Bian and D. Tao, "Constrained empirical risk minimization framework for distance metric learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1194–1205, 2012.

[156] D. Kedem, S. Tyree, F. Sha, G. R. Lanckriet, and K. Q. Weinberger, "Non-linear metric learning," in *Advances in Neural Information Processing Systems*, pp. 2573–2581, 2012.

[157] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems*, pp. 513–520, 2005.

[158] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *Proceedings of the 21th International Conference on Machine Learning*, pp. 94–101, 2004.

[159] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[160] A. Evgeniou and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems*, pp. 41–48, 2010.

[161] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1738–1754, 2012.

[162] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 252–264, 2015.

[163] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212, Springer, 2011.

[164] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

[165] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[166] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, 2014.

[167] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Artificial Neural Networks—ICANN'97*, pp. 583–588, Springer, 1997.

[168] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the tenth National Conference on Artificial Intelligence (AAAI–92)*, vol. 2, pp. 129–134, 1992.

[169] X. Chen and M. Wasikowski, "Fast: a ROC-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 124–132, ACM, 2008.

[170] K. Murphy, B. van Ginneken, A. M. Schilham, B. De Hoop, H. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," *Medical Image Analysis*, vol. 13, no. 5, pp. 757–770, 2009.

[171] Y. Ou, D. Shen, J. Zeng, L. Sun, J. Moul, and C. Davatzikos, "Sampling the spatial patterns of cancer: Optimized biopsy procedures for estimating prostate cancer volume and gleason score," *Medical Image Analysis*, vol. 13, no. 4, pp. 609–620, 2009.

[172] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, no. 4, pp. 622–639, 2011.

[173] L. Wang, "Feature selection with kernel class separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1534–1546, 2008.

[174] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu, "Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 256–269, 2011.

[175] C. Lian, S. Ruan, T. Denoux, and P. Vera, "Outcome prediction in tumour therapy based on Dempster-Shafer theory," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pp. 63–66, IEEE, 2015.

[176] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection," *Medical Image Analysis*, vol. 8, no. 3, pp. 275–283, 2004.

[177] G. Wang, S. Zhang, H. Xie, D. N. Metaxas, and L. Gu, "A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning," *Medical Image Analysis*, vol. 19, no. 1, pp. 176–186, 2015.

[178] T. D. Barwick, A. Taylor, and A. Rockall, "Functional imaging to predict tumor response in locally advanced cervical cancer," *Current Oncology Reports*, vol. 15, no. 6, pp. 549–558, 2013.

[179] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, pp. 321–357, 2002.

[180] J. Calais, S. Thureau, B. Dubray, R. Modzelewski, L. Thiberville, I. Gardin, and P. Vera, "Areas of high 18F-FDG uptake on preradiotherapy PET/CT identify preferential sites of local relapse after chemoradiotherapy for non–small cell lung cancer," *Journal of Nuclear Medicine*, vol. 56, no. 2, pp. 196–203, 2015.

[181] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.

[182] B. Efron and R. Tibshirani, "Improvements on cross-validation: the 632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[183] S. Vauclin, K. Doyeux, S. Hapdey, A. Edet-Sanson, P. Vera, and I. Gardin, "Development of a generic thresholding algorithm for the delineation of 18FDG-PET-positive tissue: application to the comparison of three thresholding models," *Physics in Medicine and Biology*, vol. 54, no. 22, p. 6901, 2009.

[184] W. Mu, Z. Chen, W. Shen, F. Yang, Y. Liang, R. Dai, N. Wu, and J. Tian, "A segmentation algorithm for quantitative analysis of heterogeneous tumors of the cervix with 18 F-FDG PET/CT," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2465–2479, 2015.

[185] M. Hatt, L. Rest, C. Cheze, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive bayesian segmentation approach for volume determination in PET," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 881–893, 2009.

[186] B. Foster, U. Bagci, Z. Xu, B. Dey, B. Luna, W. Bishai, S. Jain, and D. J. Mollura, "Segmentation of PET images for computer-aided functional quantification of tuberculosis in small animal models," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 711–724, 2014.

[187] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59–68, ACM, 2004.

[188] J. Jiang *et al.*, "Unsupervised metric learning by self-smoothing operator," in *2011 International Conference on Computer Vision*, pp. 794–801, IEEE, 2011.

[189] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[190] A. Roche, D. Ribes, M. Bach-Cuadra, and G. Krüger, "On the convergence of EM-like algorithms for image segmentation using Markov random fields," *Medical Image Analysis*, vol. 15, no. 6, pp. 830–839, 2011.

[191] A.-L. Jousselme, D. Grenier, and É. Bossé, "A new distance between two bodies of evidence," *Information Fusion*, vol. 2, no. 2, pp. 91–101, 2001.

[192] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Mathematical Programming*, vol. 107, no. 3, pp. 391–408, 2006.

[193] L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 780–795, 1999.

[194] C. Greco, K. Rosenzweig, G. L. Cascini, and O. Tamburrini, "Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC)," *Lung Cancer*, vol. 57, no. 2, pp. 125–134, 2007.

Information Fusion and Decision-Making Using Belief Functions: Application to Therapeutic Monitoring of Cancer

Radiation therapy is one of the most principal options used in the treatment of malignant tumors. To enhance its effectiveness, two critical issues should be carefully dealt with, i.e., reliably predicting therapy outcomes to adapt undergoing treatment planning for individual patients, and accurately segmenting tumor volumes to maximize radiation delivery in tumor tissues while minimize side effects in adjacent organs at risk. Positron emission tomography with radioactive tracer fluorine-18 fluorodeoxyglucose (FDG-PET) can non-invasively provide significant information of the functional activities of tumor cells.

In this thesis, the goal of our study consists of two parts: 1) to propose reliable therapy outcome prediction system using primarily features extracted from FDG-PET images; 2) to propose automatic and accurate algorithms for tumor segmentation in PET and PET-CT images. The theory of belief functions is adopted in our study to model and reason with uncertain and imprecise knowledge quantified from noisy and blurring PET images. In the framework of belief functions, a sparse feature selection method and a low-rank metric learning method are proposed to improve the classification accuracy of the evidential K-nearest neighbor classifier learnt by high-dimensional data that contain unreliable features. Based on the above two theoretical studies, a robust prediction system is then proposed, in which the small-sized and imbalanced nature of clinical data is effectively tackled. To automatically delineate tumors in PET images, an unsupervised 3-D segmentation based on evidential clustering using the theory of belief functions and spatial information is proposed. This mono-modality segmentation method is then extended to co-segment tumor in PET-CT images, considering that these two distinct modalities contain complementary information to further improve the accuracy. All proposed methods have been performed on clinical data, giving better results comparing to the state of the art ones.

Keywords: Theory of belief functions, Feature selection, Distance metric learning, Data classification, Data clustering, Cancer therapy outcome prediction, Automatic tumor segmentation, PET/CT imaging

Fusion de l'information et la prise de décisions à l'aide des fonctions de croyance : application au suivi thérapeutique du cancer

La radiothérapie est une des méthodes principales utilisée dans le traitement thérapeutique des tumeurs malignes. Pour améliorer son efficacité, deux problèmes essentiels doivent être soigneusement traités : la prédication fiable des résultats thérapeutiques et la segmentation précise des volumes tumoraux. La tomographie d'émission de positions au traceur Fluoro-18-déoxy-glucose (FDG-TEP) peut fournir de manière non invasive des informations significatives sur les activités fonctionnelles des cellules tumorales.

Les objectifs de cette thèse sont de proposer : 1) des systèmes fiables pour prédire les résultats du traitement contre le cancer en utilisant principalement des caractéristiques extraites des images FDG-TEP; 2) des algorithmes automatiques pour la segmentation de tumeurs de manière précise en TEP et TEP-TDM. La théorie des fonctions de croyance est choisie dans notre étude pour modéliser et raisonner des connaissances incertaines et imprécises pour des images TEP qui sont bruitées et floues. Dans le cadre des fonctions de croyance, nous proposons une méthode de sélection de caractéristiques de manière parcimonieuse et une méthode d'apprentissage de métriques permettant de rendre les classes bien séparées dans l'espace caractéristique afin d'améliorer la précision de classification du classificateur EK-NN. Basées sur ces deux études théoriques, un système robuste de prédiction est proposé, dans lequel le problème d'apprentissage pour des données de petite taille et déséquilibrées est traité de manière efficace. Pour segmenter automatiquement les tumeurs en TEP, une méthode 3-D non supervisée basée sur le regroupement évidentiel (evidential clustering) et l'information spatiale est proposée. Cette méthode de segmentation mono-modalité est ensuite étendue à la co-segmentation dans des images TEP-TDM, en considérant que ces deux modalités distinctes contiennent des informations complémentaires pour améliorer la précision. Toutes les méthodes proposées ont été testées sur des données cliniques, montrant leurs meilleures performances par rapport aux méthodes de l'état de l'art.

Mots-clés : La théorie de fonctions de croyance, Sélection des caractéristiques, Apprentissage de métriques, Classification des données, Clustering des données, Prédiction, Radiothérapie, Segmentation de tumeurs automatique, Imagerie TEP/TDM