Selecting Radiomic Features from FDG-PET Images for Cancer Treatment Outcome Prediction

Chunfeng Lian[a,b], Su Ruan[b], Thierry Denœux[a], Fabrice Jardin[b,d], Pierre Vera[b,c]

[a] *Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France*
[b] *Université de Rouen, QuantIF, EA 4108 LITIS, France*
[c] *Centre Henri-Becquerel, Department of Nuclear Medicine, France*
[d] *Centre Henri-Becquerel, Department of Hematology, France*

Abstract

As a vital task in cancer therapy, accurately predicting the treatment outcome is valuable for tailoring and adapting a treatment planning. To this end, multi-sources of information (radiomics, clinical characteristics, genomic expressions, etc) gathered before and during treatment are potentially profitable. In this paper, we propose such a prediction system primarily using radiomic features (e.g., texture features) extracted from FDG-PET images. The proposed system includes a feature selection method based on Dempster-Shafer theory, a powerful tool to deal with uncertain and imprecise information. It aims to improve the prediction accuracy, and reduce the imprecision and overlaps between different classes (treatment outcomes) in a selected feature subspace. Considering that training samples are often small-sized and imbalanced in our applications, a data balancing procedure and specified prior knowledge are taken into account to improve the reliability of the selected feature subsets. 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.

Keywords: Dempster-Shafer theory, feature selection, imbalanced learning, outcome prediction, cancer, PET images

1. Introduction

Accurate outcome prediction prior to or even during cancer therapy is of great clinical value. It benefits the adaptation of more effective treatment planning for individual patient. With the advances in medical imaging technology, radiomics, referring to the extraction and analysis of a large amount of quantitative image features, provide an unprecedented opportunity to improve personalized treatment assessment (Aerts et al., 2014). Positron emission tomography (PET), with the radio-tracer fluoro-2-deoxy-D-glucose (FDG), is one of the important and advanced imaging tools for diagnosis, staging , and restaging of cancers. According to practice guidelines presented by the Society of Nuclear Medicine and Molecular Imaging (SNMMI) [1], FDG-PET or FDG-PET/CT is now playing an essential role in clinical oncology, such as initial staging and gross tumor volume delineation for lung cancer patients receiving radiotherapy ; initial staging and restaging of esophageal cancer ; and routine pre-treatment staging and restaging of patients with Hodgkin lymphoma and many subtypes of non-Hodgkin lymphoma, etc.

Apart from diagnosis and staging, the functional information provided by FDG-PET has also emerged to be predictive of the pathologic response of a treatment in some types of cancers, such as lung tumor, esophageal tumor (Tan et al., 2013) and cervix tumor (Barwick et al., 2013). For this application, variety radiomic features are well-explored on FDG-PET (Cook et al., 2014), which include standardized uptake values (SUVs), e.g., SUV_{max}, SUV_{peak} and SUV_{mean}, to describe metabolic uptakes in a region of interest (ROI), and metabolic tumor volume (MTV) and total lesion glycolysis (TLG) to describe metabolic tumor burdens. Apart from SUV-based features, some complementary characterization of PET images, like texture analysis (Tixier et al., 2011) and shape analysis (El Naqa et al., 2009), may also provide supplementary knowledge associated with the treatment outcome. Although the quantification of these

1. http ://www.snmmi.org/ClinicalPractice/

radiomic features, as well as the calculation of their temporal changes during the treatment, have been claimed to have the discriminative power (Aerts et al., 2014), the solid application is still hampered by some practical difficulties :

First, abounding features (e.g., radiomics and clinical characteristics) can be collected for outcome prediction, but without any consensus to determine the most discriminative factors among them. Thus, finding information regarding the most predictive features could be interesting from the point of clinicians.

Second, comparing to a relatively large amount of input features, only a limited number of observations (small data size is often encountered in the medical domain) are available for constructing a prediction system. A high dimensional feature space may increase the complexity of the learning models, thus leading to high risk of over-fitting on the small-sized learning set.

Third, it often happens that some of the input features are irrelevant with the outcome label. Moreover, badly defined features sometimes may even degrade the performance of a prediction model.

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 (Guyon and Elisseeff, 2003). 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) (Kira and Rendell, 1992) 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) (Chen and Wasikowski, 2008), 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 (Guyon and Elisseeff, 2003).

Different from ranking features, multivariate methods evaluate a subset of features ensemble. Sequential Forward Selection (SFS) and Sequential Forward Floating Selection (SFFS) (Pudil et al., 1994) are two classical subset selec-

tion 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 (Murphy et al., 2009). It in some sense reduces the nesting problem of SFS, but still has the risk to be sub-optimal with limited learning instances (Mi et al., 2015). 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 (Mi et al., 2015). 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) (Guyon et al., 2002). The variants of this method have been successfully applied for prostate cancer volume estimation (Ou et al., 2009) and deformable registration in medical imaging (Ou et al., 2011). 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 (Wang, 2008). As a robust method, KCS has found promising application for tumor delineation in multi-spectral MRI images (Zhang et al., 2011). 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 (Somol and Novovicova, 2010), 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, (Lian et al., 2015a) proposed a robust subset selection method, namely Evidential Feature Selection (EFS), based on the Dempster-Shafer Theory (DST) (Shafer, 1976), a powerful tool for modeling and reasoning with uncertain and/or imprecise information. 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 (Brooks and Grigsby, 2014), 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 (Lian et al., 2015b). However, the imbalanced learning problem in feature selection (another important issue of medical data) is still left unsolved for this method.

In this paper, 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 (Denœux, 1995), a stable classification method based on DST, is used with selected feature subsets to output prediction results.

The rest of this paper is organized as follows. The background on DST and the original EFS is recalled in Section 2. Then, an improved EFS with prior knowledge and data balancing is introduced in Section 3. The proposed

method is evaluated by three clinical datasets described in Section 4, and the experimental results are summarized in Section 5. Some discussions and the conclusion are presented in Section 6 and Section 7, respectively.

2. Background

The necessary background on DST and the original EFS is briefly reviewed in Sections 2.1 and 2.2, respectively.

2.1. Dempster-Shafer Theory

DST is also known as the theory of belief functions or Evidence theory. As an extension of probability theory and the set-membership approach, DST has shown remarkable applications in divers fields, such as medical image processing (Bloch, 1996; Lelandais et al., 2014; Makni et al., 2014), statistical machine learning (Zhu and Basir, 2005; Denœux and Smets, 2006; Masson and Denœux, 2008; Liu et al., 2015), and computer vision (Xu et al., 2014; Wang et al., 2014) etc. DST consists of two main components, i.e., the quantification of a piece of evidence and the combination of different items of evidence.

2.1.1. Evidence Quantification

DST is a formal framework for reasoning under uncertainty based on the modeling of evidence (Shafer, 1976). 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

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

Each number $m(A)$ denotes a *degree of belief* attached to the hypothesis that $\omega \in A$. Any subset A with $m(A) > 0$ is called a *focal element* of mass function m. Function m is said to be normalized if $m(\emptyset) = 0$. Corresponding to a normalized

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

$$Bel(A) = \sum_{B \subseteq A} m(B); \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2}$$

Quantity $Bel(A)$ can be interpreted as the degree to which the evidence supports the hypothesis $\omega \in A$, while $Pl(A)$ can be interpreted as the degree to which the evidence is not contradictory to that hypothesis. 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.

2.1.2. Evidence Combination

In DST, beliefs are refined by aggregating different items of evidence. *Dempster's rule of combination* (Shafer, 1976), as well as its unnormalized version, i.e., the *conjunctive combination rule* defined in the Transferable Belief Model (TBM) (Smets and Kennes, 1994), are basic mechanisms for evidence fusion.

Let m_1 and m_2 be two mass functions derived from two independent items of evidence. They can be fused via the TBM conjunctive rule to induce a new mass function $m_{1\bigcirc\!\!\!\cap 2}$ defined as

$$m_{1\bigcirc\!\!\!\cap 2}(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{3}$$

This new mass function reduces uncertainty and imprecision via transferring masses of belief to conjunctions of the focal elements. Quantity $m_{1\bigcirc\!\!\!\cap 2}(\emptyset)$ measures the *degree of conflict* between evidence m_1 and m_2. If $m_{1\bigcirc\!\!\!\cap 2}(\emptyset) < 1$, the new mass function obtained by Dempster's rule can be represented as

$$m_{1 \bigoplus 2}(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \frac{m_{1\bigcirc\!\!\!\cap 2}(A)}{1 - m_{1\bigcirc\!\!\!\cap 2}(\emptyset)} & \text{otherwise.} \end{cases} \tag{4}$$

As can be seen, Dempster's rule normalizes the conflict obtained by the TBM conjunctive rule. Both the TBM conjunctive rule and Dempster's rule are commutative and associative. They can be easily generalized to combine N (≥ 2) independent sources of information.

2.2. Evidential Feature Selection

Let $\{(X_i, Y_i)|i = 1, \cdots, N\}$ be a collection of N training pairs, in which $X_i = [x_{i,1}, \cdots, x_{i,V}]^T$ is the ith training instance with V features, and $Y_i \in \{\omega_1, \cdots, \omega_c\}$ is the corresponding class label.

EFS (Lian et al., 2015a) searches for a qualified feature subset according to three requirements : first, high classification accuracy ; second, low imprecision and uncertainty, namely small overlaps between different classes in the output feature space ; third, sparsity to reduce the risk of over-fitting. To learn such a feature subset, EFS uses a weighted Euclidian distance measure to represent the dissimilarity between any two training instances. Hence, the dissimilarity between X_i and X_j is

$$d_{i,j} = \sqrt{\sum_{p=1}^{V} \lambda_p {d_{ij,p}}^2}, \tag{5}$$

where $d_{ij,p} = |x_{i,p} - x_{j,p}|$ represents the difference between the pth dimension of X_i and X_j. Features are selected via changing the value of the *binary vector* $\Lambda = [\lambda_1, \ldots, \lambda_V]^T$. As the result, the pth dimension of the input feature space is selected when $\lambda_p = 1$, while eliminated when $\lambda_p = 0$.

We orderly regard each training instance X_i as a query object. Then, other samples in the training pool can be considered as independent items of evidence that support different hypotheses regarding the class membership of X_i. The evidence offered by the training sample $(X_j, Y_j = \omega_q)$ is partially reliable, and can be modeled by a mass function

$$\begin{cases} m_{i,j}(\{\omega_q\}) & = e^{-\gamma_q d_{i,j}^2}, \\ m_{i,j}(\Omega) & = 1 - e^{-\gamma_q d_{i,j}^2}, \end{cases} \tag{6}$$

where $d_{i,j}$ is the distance between X_i and X_j that measured by (5). Positive parameters $\gamma = [\gamma_1, \ldots, \gamma_c]^T$ are set as the inverse of the mean distance between training instances from the same class.

After obtaining all the independent mass functions for X_i, they can be further fused by a mixed combination rule, called Dempster+Yager rule (Lian et al., 2015a), so as to obtain a global one describing the class membership of X_i. This

rule consists of two main steps : first, using Dempster's rule to combine mass functions originated from the same class, and discounting the resulting mass function according to the number of instances in this class ; second, combining the discounted mass functions originated from different classes via the Yager's rule (Yager, 1987) to output the global mass function. This combination procedure integrates the advantages of Dempster's and Yager's rules, thus could robustly represent all imprecision and uncertainty of the training data on the whole frame of discernment (i.e., Ω).

Finally, based on the global mass functions for all training samples, a loss function with respect to the binary vector $\Lambda = [\lambda_1, \ldots, \lambda_V]^T$ is constructed for feature selection,

$$\arg\min_{\Lambda} \frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{c} \{Pl_i(\{\omega_q\}) - t_{i,q}\}^2 + \frac{1}{N} \sum_{i=1}^{N} m_i(\Omega) + \beta ||\Lambda||_0, \qquad (7)$$

where m_i and Pl_i, concerning Λ, are the global mass function and the corresponding plausibility function of the training instance X_i. The first term of (7) is a mean squared error measure corresponding to the first requirement of EFS (namely high classification accuracy). Binary vector t_i is the class label indicator, with $t_{i,q} = \delta_{i,q}$ if and only if $Y_i = \omega_q$. The second term of (7) penalizes feature subsets that lead to high overlaps between different classes, thus corresponding to the second requirement of EFS. The last term , which is an approximation of the l_0-norm of Λ, forces the selected features to be sparse, thus realizing the last requirement of EFS. Parameter β controls the influence of this sparsity penalty.

The mixed combination rule used in the original EFS can lead to robust quantification of data uncertainty and imprecision. However, since a discounting procedure is included, additional parameters increase method's complexity. To cope with this problem, we propose a new method described in the next section.

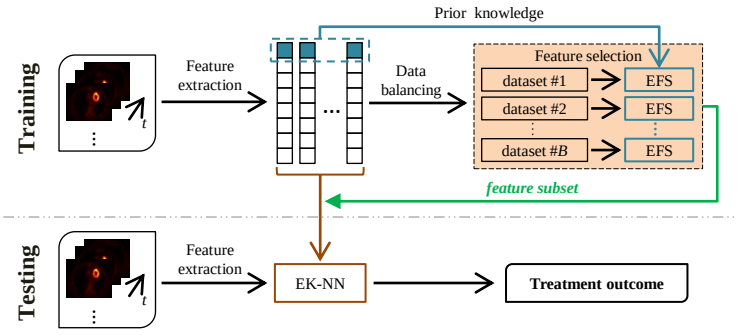


FIGURE 1: Protocol of the prediction system.

3. 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 paper.

3.1. Main Protocol

The rough protocol of the prediction system is shown in Figure 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 feture 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.

3.2. 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 (Lemarignier et al., 2014). Three types of PET imaging features are quantified, namely SUV-based features, texture features, and the temporal changes of these two types of features.

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 the Appendix (Table A.7).

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 (Tixier et al., 2011), Gray Level Size Zone Matrix (GLSZM) (Thibault et al., 2014) 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 the Appendix (Table A.8).

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.

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.

3.3. Improved EFS

To reduce the complexity of the original EFS, 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. As discussed in Section 2.2, the evidence offered by each training instance X_j ($\neq i$) can be quantified as a mass function using (6) and (5). 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. Correspondingly, $\{m_{i,i_1}, \ldots, m_{i,i_K}\}$ 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 (3) 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

$$\begin{cases} m_i^{\Theta_q}(\{\omega_q\}) & = 1 - \prod_{X_{i_p} \in \Theta_q}^{p=1,\ldots,K} \left(1 - e^{-\gamma_q d_{i,i_p}^2}\right), \\ m_i^{\Theta_q}(\Omega) & = \prod_{X_{i_p} \in \Theta_q}^{p=1,\ldots,K} \left(1 - e^{-\gamma_q d_{i,i_p}^2}\right); \end{cases} \tag{8}$$

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,

$$\begin{cases} M_i(\{\omega_1\}) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\Omega), \\ M_i(\{\omega_2\}) & = m_i^{\Theta_2}(\{\omega_2\}) \cdot m_i^{\Theta_1}(\Omega), \\ M_i(\Omega) & = m_i^{\Theta_1}(\Omega) \cdot m_i^{\Theta_2}(\Omega), \\ M_i(\emptyset) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\{\omega_2\}). \end{cases} \tag{9}$$

Based on (5) to (9), 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 Section 2.2, the new loss function with respect to Λ can be defined as

$$L(\Lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{2} \{M_i(\{\omega_q\}) - t_{i,q}\}^2 + \frac{1}{N} \sum_{i=1}^{N} \{M_i(\emptyset)^2 + M_i(\Omega)^2\} + \beta ||\Lambda||_0. \tag{10}$$

In (10), 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 (10) is integer constrained (vector Λ should be binary), an integer Genetic Algorithm (GA), namely the MI-LXPM (Deep et al., 2009), 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 optima 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.

3.4. Prior Knowledge

Prior information, such as spatial constraints (Prastawa et al., 2004), shape prior (Wang et al., 2015) 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 (Tan et al., 2013; Barwick et al., 2013), we incorporate this important information into EFS as a predefined constraint.

More specifically, a feature ranking method, namely RELIEF (Kira and Rendell, 1992), 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

$$S(\tilde{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{k}\sum_{j=1}^{k} diff(\tilde{f}, X_i, miss_j^i) - \frac{1}{k}\sum_{j=1}^{k} diff(\tilde{f}, X_i, hit_j^i)\right), \quad (11)$$

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 (11) was 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 (10). This added constraint drives EFS into a confined searching space. It ensures more robust feature selection, thus increasing the reliability of the prediction system.

3.5. 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 (Pudil et al., 1994) 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 (Chawla et al., 2002). It has shown plenty of successes in many applications, and its variants, such as ADAptive SYNthetic sampling (ADASYN) (He et al., 2009), 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).

Algorithm 1: ADASYN-based balancing for feature selection (He et al., 2009)

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.

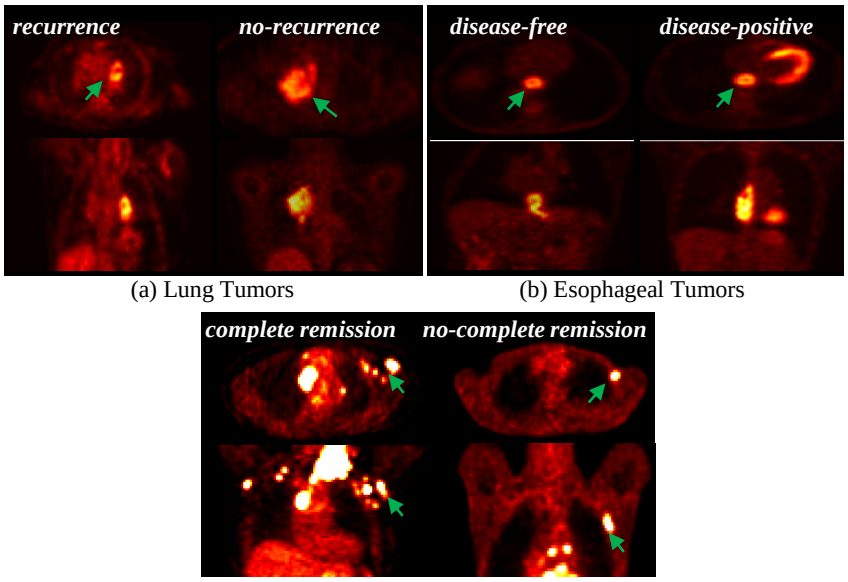
Set the number of synthetic minority class instances as $n_{syn} = n_{maj} - n_{min}$.
for *each sample* X_j *with* $Y_j = \omega_1$ **do**
    Find k nearest neighbors of X_j in the training pool.
    Calculate the parameter r_j for X_j as $r_j = \Delta_j/k$, where Δ_j is the number of nearest neighbors of X_j that belong to the majority class.
end
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 $S_l^j = X_j + \delta \times (X_r - X_j)$.
    end
end

Given an imbalanced training dataset, ADASYN outputs an balanced training dataset via the procedure summarized in Algorithm 1. 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 1 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.

3.6. 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 (Denœux, 1995)

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 (Zouhal and Denœux, 1998). 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 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.

4. Clinical Datasets

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

1) Lung Cancer 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 (Calais et al., 2015). 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 (Lemarignier et al., 2014). 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 (Lanic et al., 2012). 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 2.

Feature Description. As discussed in Section 3.2, 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 (Lanic et al., 2012). The three clinical datasets are briefly summarized in Table 1, 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 1: Description of the three clinical datasets.

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

5. 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 (Lemarignier et al., 2014)).

5.1. 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 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, which ensures low bias and variance estimation (Efron and Tibshirani, 1997).

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 (Somol and Novovicova, 2010). 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 average Accuracy was adopted, which is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{12}$$

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, the simple 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

$$TP_{rate} = \frac{TP}{TP + FN}; \quad FP_{rate} = \frac{FP}{TN + FP}. \tag{13}$$

Based on the ROC curve, 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. 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 (Mi et al., 2015).
- 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.

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 2, where the results for all the input features (the SVM classifier was used) are also presented as the baselines for comparison. From Table 2 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

TABLE 2: Feature selection performance evaluated by the LOOCV. EFS represents our previous work (Lian et al., 2015a), 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	—	0.64	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

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 (Lian et al., 2015a) 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.

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 3. Consistent with the results presented in Table 2, the robustness of the proposed ***i***EFS that evaluated by the bootstrapping was still better than other methods

TABLE 3: Feature selection performance evaluated by the .632+ Bootstrapping. EFS represents our previous work (Lian et al., 2015a), 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

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 3 with that in Table 2, 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.

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

Table 4: 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: 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

in Table 4 to Table 6, respectively. For the lung tumor (Table 4), the SUV_{max} during the fith week of RT (PET_2) has also been proven to have significant predictive power in the clinical study (Calais et al., 2015) ; for the esophageal tumor (Table 5), the role of the TLG at tumor staging (PET_0) has been clinically validated in (Lemarignier et al., 2014) ; and for the lymph tumor (Table 6), 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 (Lanic et al., 2012).

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.

5.2. 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 (pre-

TABLE 6: 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.

dictors) which have been clinically validated and discussed in the last part of Section 5.1. The average AUC with respect to different K is shown in Figure 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.

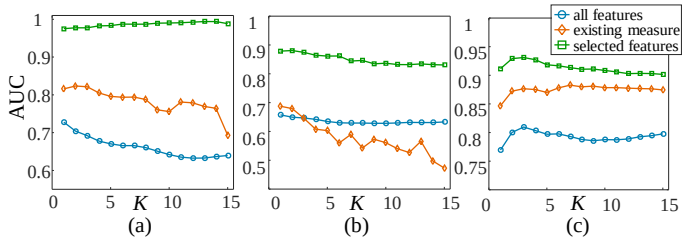


FIGURE 3: Prediction performance of the EK-NN classifier with respect to different K : (a) lung tumor dataset, (b) esophageal tumor dataset, and (c) lymph tumor dataset. "all features", "selected features", and "existing measure" denote the results obtained by the input features, the selected feature subset and the predictor that has been clinically proven, respectively.

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 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.

6. Discussion

Influence of imbalance level. According to the analysis in Section 5.1, 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.

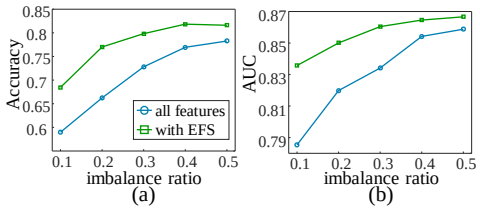


FIGURE 4: (a) Accuracy, and (b) AUC for the synthetic dataset.

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 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.

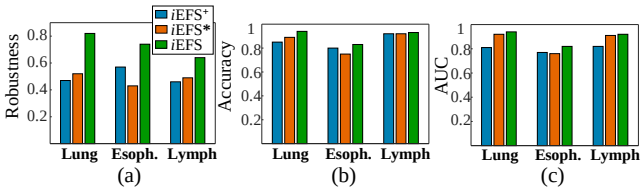


FIGURE 5: (a) Subset robustness, (b) Accuracy, and (c) AUC that evaluated by the .632+ Bootstrapping for the improved EFS without data balancing ($\boldsymbol{i}$EFS^{+}), the improved EFS without prior knowledge ($\boldsymbol{i}$EFS*), and the improved EFS ($\boldsymbol{i}$EFS), respectively.

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, in which $\boldsymbol{i}$EFS denotes the improved EFS used in our prediction system; while, $\boldsymbol{i}$EFS^{+} and $\boldsymbol{i}$EFS* denote $\boldsymbol{i}$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.

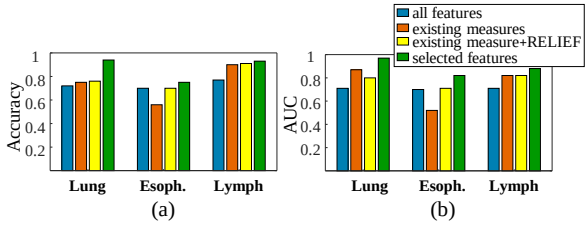


FIGURE 6: (a) Accuracy and (b) AUC of the logistic regression method that evaluated by the .632+ Bootstrapping. The selected features were compared with all the input features, the clinically validated predictors (i.e., existing measures), and the clinically validated predictors joint with features selected by the classical RELIEF (i.e., existing measure+RELIEF).

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 4 to Table 6. 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 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.

Multi-class problems. Apart from the binary-class examples discussed in this paper, the proposed method can also be easily generalized to handle the multi-class ($c \geq 2$) problems. To this end, we need to replace (9) with

$$\begin{cases} M_i(\{\omega_q\}) & = m_i^{\Theta_q}(\{\omega_q\}) \prod_{p \neq q}^{c} m_i^{\Theta_p}(\Omega), \forall q \in \{1, \ldots, c\} \\ M_i(\Omega) & = \prod_{q=1}^{c} m_i^{\Theta_q}(\Omega) \\ M_i(\emptyset) & = 1 - \sum_{q=1}^{c} M_i(\{\omega_q\}) - M_i(\Omega) \end{cases}, \tag{14}$$

and change the first term of (10) as $\frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{c} \{M_i(\{\omega_q\}) - t_{i,q}\}^2$.

7. Conclusion

A new framework for PET imaging based cancer treatment outcome prediction has been proposed in this paper. 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 (Lian et al., 2015a), 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.

In the future, to further improve the reliability of our prediction system, we plan to include more radiomic features extracted from other image modalities, such as CT, MRI and multi-tracer PETs. In addition, to tackle the imbalanced learning problems, other data balancing or cost-sensitive learning methods should be studied and compared with the method that has been used in this paper.

Appendix A. Radiomic Features Extracted from PET Imaging

TABLE A.7: 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

TABLE A.8: Definition of GLSZM-based features (Thibault et al., 2014). 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.

References

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

Barwick, T.D., Taylor, A., Rockall, A., 2013. Functional imaging to predict tumor response in locally advanced cervical cancer. Current Oncology Reports 15, 549–558.

Bloch, I., 1996. Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. Pattern Recognition Letters 17, 905–919.

Brooks, F.J., Grigsby, P.W., 2014. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. Journal of Nuclear Medicine 55, 37–42.

Calais, J., Thureau, S., Dubray, B., Modzelewski, R., Thiberville, L., Gardin, I., Vera, P., 2015. 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 56, 196–203.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE : synthetic minority over-sampling technique. Journal of Artificial Intelligence Research , 321–357.

Chen, X., Wasikowski, M., 2008. 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, ACM. pp. 124–132.

Cook, G.J., Siddique, M., Taylor, B.P., Yip, C., Chicklore, S., Goh, V., 2014. Radiomics in PET : principles and applications. Clinical and Translational Imaging 2, 269–276.

Deep, K., Singh, K.P., Kansal, M., Mohan, C., 2009. A real coded genetic algorithm for solving integer and mixed integer optimization problems. Applied Mathematics and Computation 212, 505–518.

Denœux, T., 1995. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25, 804–813.

Denœux, T., Smets, P., 2006. Classification using belief functions : relationship between case-based and model-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics 36, 1395–1406.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation : the 632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

El Naqa, I., Grigsby, P., Apte, A., Kidd, E., Donnelly, E., Khullar, D., Chaudhari, S., Yang, D., Schmitt, M., Laforest, R., et al., 2009. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognition 42, 1162–1171.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning 46, 389–422.

He, H., Garcia, E., et al., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284.

Kira, K., Rendell, L.A., 1992. The feature selection problem : Traditional methods and a new algorithm, in : Proceedings of the tenth National Conference on Artificial Intelligence (AAAI–92), pp. 129–134.

Lanic, H., Mareschal, S., Mechken, F., Picquenot, J.M., Cornic, M., Maingonnat, C., Bertrand, P., Clatot, F., Bohers, E., Stamatoullas, A., et al., 2012. 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 53, 34–42.

Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I., 2014. Fusion of multi-tracer PET images for dose painting. Medical Image Analysis 18, 1247–1259.

Lemarignier, C., Di Fiore, F., Marre, C., Hapdey, S., Modzelewski, R., Gouel, P., Michel, P., Dubray, B., Vera, P., 2014. 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 41, 2008–2016.

Lian, C., Ruan, S., Denœux, T., 2015a. An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognition 48, 2318–2327.

Lian, C., Ruan, S., Denoux, T., Vera, P., 2015b. Outcome prediction in tumour therapy based on Dempster-Shafer theory, in : Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE. pp. 63–66.

Liu, Z.G., Pan, Q., Dezert, J., Mercier, G., 2015. Credal c-means clustering method based on belief functions. Knowledge-Based Systems 74, 119–132.

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

Masson, M.H., Denœux, T., 2008. ECM : An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.

Mi, H., Petitjean, C., Dubray, B., Vera, P., Ruan, S., 2015. Robust feature selection to predict tumor treatment outcome. Artificial Intelligence in Medicine 64, 195–204.

Murphy, K., van Ginneken, B., Schilham, A.M., De Hoop, B., Gietema, H., Prokop, M., 2009. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Medical Image Analysis 13, 757–770.

Ou, Y., Shen, D., Zeng, J., Sun, L., Moul, J., Davatzikos, C., 2009. Sampling the spatial patterns of cancer : Optimized biopsy procedures for estimating prostate cancer volume and gleason score. Medical Image Analysis 13, 609–620.

Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C., 2011. DRAMMS : Deformable registration via attribute matching and mutual-saliency weighting. Medical Image Analysis 15, 622–639.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Medical Image Analysis 8, 275–283.

Pudil, P., Novovičová, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15, 1119–1125.

Shafer, G., 1976. A mathematical theory of evidence. Princeton University Press.

Smets, P., Kennes, R., 1994. The transferable belief model. Artificial Intelligence 66, 191–234.

Somol, P., Novovicova, J., 2010. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1921–1939.

Tan, S., Kligerman, S., Chen, W., Lu, M., Kim, G., Feigenberg, S., D'Souza, W.D., Suntharalingam, M., Lu, W., 2013. 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 85, 1375–1382.

Thibault, G., Angulo, J., Meyer, F., 2014. Advanced statistical matrices for texture characterization : application to cell classification. IEEE Transactions on Biomedical Engineering 61, 630–637.

Tixier, F., Le Rest, C.C., Hatt, M., Albarghach, N., Pradier, O., Metges, J.P., Corcos, L., Visvikis, D., 2011. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. Journal of Nuclear Medicine 52, 369–378.

Wang, F., Miron, A., Ainouz, S., Bensrhair, A., 2014. Post-aggregation stereo matching method using Dempster-Shafer theory, in : 2014 IEEE International Conference on Image Processing (ICIP 2014), IEEE. pp. 3783–3787.

Wang, G., Zhang, S., Xie, H., Metaxas, D.N., Gu, L., 2015. A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning. Medical Image Analysis 19, 176–186.

Wang, L., 2008. Feature selection with kernel class separability. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 1534–1546.

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

Yager, R.R., 1987. On the Dempster-Shafer framework and new combination rules. Information Sciences 41, 93–137.

Zhang, N., Ruan, S., Lebonvallet, S., Liao, Q., Zhu, Y., 2011. Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. Computer Vision and Image Understanding 115, 256–269.

Zhu, H., Basir, O., 2005. An adaptive fuzzy evidential nearest neighbor formulation for classifying remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 43, 1874–1889.

Zouhal, L.M., Denœux, T., 1998. An evidence-theoretic k-NN rule with parameter optimization. IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews 28, 263–271.