Reliable gene mutation prediction in clear cell renal cell carcinoma through multi-classifier multi-objective radiogenomics model

Xi Chen1, Zhiguo Zhou2, Raquibul Hannan3, Kimberly Thomas3, Ivan Pedrosa3, Payal Kapur2,4, James Brugarolas2,5, Xuanqin Mou1 and Jing Wang6

1 School of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, People’s Republic of China
2 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
3 Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
4 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
5 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
6 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
7 Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America

E-mail: Zhiguo.Zhou@UTSouthwestern.edu and Jing.Wang@UTSouthwestern.edu

Keywords: radiogenomics, multi-objective optimization, multi-classifier, outcome prediction, evidential reasoning

Abstract

Genetic studies have identified associations between gene mutations and clear cell renal cell carcinoma (ccRCC). Since the complete gene mutational landscape cannot be characterized through biopsy and sequencing assays for each patient, non-invasive tools are needed to determine the mutation status for tumors. Radiogenomics may be an attractive alternative tool to identify disease genomics by analyzing amounts of features extracted from medical images. Most current radiogenomics predictive models are built based on a single classifier and trained through a single objective. However, since many classifiers are available, selecting an optimal model is challenging. On the other hand, a single objective may not be a good measure to guide model training. We proposed a new multi-classifier multi-objective (MCMO) radiogenomics predictive model. To obtain more reliable prediction results, similarity-based sensitivity and specificity were defined and considered as the two objective functions simultaneously during training. To take advantage of different classifiers, the evidential reasoning (ER) approach was used for fusing the output of each classifier. Additionally, a new similarity-based multi-objective optimization algorithm (SMO) was developed for training the MCMO to predict ccRCC related gene mutations (VHL, PBRM1 and BAP1) using quantitative CT features. Using the proposed MCMO model, we achieved a predictive area under the receiver operating characteristic curve (AUC) over 0.85 for VHL, PBRM1 and BAP1 genes with balanced sensitivity and specificity. Furthermore, MCMO outperformed all the individual classifiers, and yielded more reliable results than other optimization algorithms and commonly used fusion strategies.

1. Introduction

Kidney cancer, most predominantly, renal cell carcinoma (RCC) remains one of the most common renal malignancies with 63,990 new cases expected to be diagnosed and with 14,400 deaths in the United States in 2017 (Siegel et al 2017). Clear cell RCC (ccRCC) is the most abundant (~75%) subtype of RCC and the most likely to metastasize outside the kidney (Motzer et al 2002). Most cases of ccRCC present with somatic (or germ line) inactivating mutations in the von Hippel–Lindau tumor suppressor (VHL) gene, which are generally absent in other cancers (Gnarra et al 1994, Varela et al 2011, Guo et al 2012, Cancer Genome Atlas Research Network 2013). Several other mutations in genes involved in regulating chromatin states, including those in the BRCA1-associated protein 1 (BAP1), polybromo 1 (PBRM1), SET domain containing 2 (SETD2), and lysine (K)-specific demethylase 5C (KDM5C), were recently identified (Dalgliesh et al 2010, Duns et al 2010, Varela et al 2011, Peña-Llopis et al 2012). Mutations in BAP1 and SETD2 were found to be associated with advanced
stage and poor outcome (Cancer Genome Atlas Research Network 2013, Hakimi et al 2013, Kapur et al 2013). The genes mutated within a tumor can be used as biomarkers and may help with prognosis, treatment selection, and treatment response prediction. However, inter- and intra-tumoral heterogeneity in gene mutations has previously been described in ccRCC (Gerlinger et al 2012, 2014, McGranahan and Swanton 2015). As ccRCC metastasizes, additional gene mutations accumulate. Because the complete gene mutational landscape is hard to be characterized for each patient through biopsy and sequencing assays, a non-invasive tool would be useful to identify the mutations within the tumor.

Radiogenomics (Rutman and Kuo 2009, Jaffe 2012, Karlo et al 2014, Kuo and Jamshidi 2014, Shinagare et al 2015, Sala et al 2017), an integrated approach that combines radiology and genomics, is based on extracting and analyzing amounts of data from clinical information by high-throughput computing. Therefore, radiogenomics is a promising solution for predicting gene mutation in ccRCC. Contrast-enhanced computed tomography (CT) is commonly used to diagnose and characterize renal masses, monitor growth in pathologically-proven RCC undergoing active surveillance, assess RCC location and extent, and determine stage and treatment response (Stewartmerrill et al 2015, Motzer et al 2017). Furthermore, the diagnostic standard of reference has expanded to the genomic level and has led to the attempt to use imaging as a noninvasive determinant of mutational status (Reznek 2004, Kuo and Yamamoto 2011, Carles et al 2012, Powles and Albers 2012). Therefore, a CT based radiogeneomics predictive model would be helpful.

In recent years, researchers have investigated predictable and systematic associations between imaging features and underlying molecular and genomic alterations in different cancers. Yamamoto et al (2012) carried out a radiogenomic analysis of breast cancer with MRI, a novel approach that may help reveal the underlying molecular biology of breast cancers. Gevaert et al (2017) used CT image features to predict the mutation status of EGFR in non-small cell lung cancer (NSCLC). Aerts et al (2014) revealed that a prognostic radiometric feature set, capturing intra-tumor heterogeneity, is associated with underlying gene-expression patterns. One study reported associations between CT features of 58 ccRCCs and the underlying karyotype (Sauk et al 2011). Others studied the association between CT imaging features and mutational status of ccRCC (Karlo et al 2014, Shinagare et al 2015). For example, the BAP1 mutation was associated with ill-defined tumor margins and calcification (Shinagare et al 2015).

By quantitatively analyzing large amounts of information from medical images, radiogenomics holds great potential to predict gene mutation. However, several challenges need to be addressed to build an optimal predictive model. First, a single classifier is typically used to build a radiogenomics predictive model. Aerts et al (2014) used the Cox proportional hazards regression model to predict survival in patients with lung and head-and-neck cancer. Other researchers tested different types of classifiers and chose one or two ‘preferred’ ones for specific applications. Valdes et al (2016) evaluated three different classifiers, including decision trees, random forests, and RUSBoost, to predict radiation pneumonitis in patients with stage I NSCLC treated with stereotactic body radiation therapy (SBRT). Higher accuracy was achieved when the RUSBoost algorithm was used with regularization. These findings indicate how difficult it is to select a ‘preferred’ classifier for a specific application. Instead of trying to find the most suitable classifier for a particular application, a model that combines multiple classifiers can fully use information from different classifiers to improve accuracy in radiogenomics. Second, most current radiogenomics models adopt a single objective function (e.g. accuracy, AUC), which may not be a good measure for building the predictive model, especially when positive and negative cases are imbalanced. To overcome the disadvantages of using a single classifier and a single objective function, we sought to develop a multi-classifier multi-objective (MCMO) radiogenomics model to predict most mutations in most commonly mutated genes in ccRCC. In MCMO, multiple classifiers are used for building the model and a multi-objective optimization algorithm is used for training the model.

2. Materials and methods

2.1. Data
2.1.1. Patients

We conducted an institutional review board-approved, Health Insurance Portability and Accountability Act-compliant (HIPAA), retrospective study including 57 ccRCC patients from two independent cohorts. The first cohort consisted of 33 patients (median age 62 years, range 28–83) from the University of Texas Southwestern Medical Center (UTSW). The other cohort consisted of 24 patients (median age 59 years, range 26–74) from the cancer genome atlas kidney renal clear cell carcinoma (TCGA-KIRC) data collection. The TCGA-KIRC data collection is part of The Cancer Genome Atlas (TCGA), an ongoing project funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which created an atlas of genetic changes related to more than 20 tumor types, including ccRCC. Clinical, genetic, and pathological data reside in the TCGA data portal, and radiological data is stored in The Cancer Imaging Archive (TCIA). Both TCGA and TCIA are accessible for public download (Clark et al 2013, Linehan et al 2016).
All 57 patients fulfilled the following criteria: (a) histopathologic diagnosis of ccRCC and exome sequencing, including information on VHL, PBRM1, and BAP1 gene mutations, considered as frequent mutations in ccRCC; (b) availability of images from a pretreatment contrast-enhanced CT including a corticomedullary phase. For each gene, the numbers of patients with mutation and without mutation are listed (table 1).

2.1.2. CT image features

CT images from UTSW were acquired by GE LightSpeed VCT (GE Healthcare, Waukesha, WI) or TOSHIBA Aquilion ONE (Canon Medical Systems USA, Tustin, CA). CT image size was $512 \times 512$ with a pixel size of 0.7–1 mm, and slice thickness was 3 or 5 mm.

The primary tumor contour was delineated by a radiation oncologist with 4 years of experience and reviewed by a radiation oncologist with nine years of experience. Contrast enhanced CT images acquired during the corticomedullary phase were used in all 57 cases for image analysis. A region of interest (ROI) was drawn along the outer contour of the mass using the Velocity 3.2.0 software excluding adjacent tissues (e.g. renal parenchyma, peri-renal fat) (figure 1).

We resampled all images of the same slice thickness at 5 mm. We only considered primary tumors and defined 43 quantitative image features describing tumor characteristics, including 13 geometry features, nine intensity features, and 21 texture features (table 2). Geometry features describing tumor shape and size were calculated according to the actual pixel size. Features Size_X, Size_Y and Size_Z (table 2) describe the tumor size along the X, Y, and Z axes of the digital imaging and communications in medicine (DICOM) coordinate system (figure 2(a)). The shape and location of the tumors differed from patient to patient. To intuitively describe the size of the tumor, a principal component analysis (PCA) was applied to the tumor contour points to transform the data into a new coordinate system. Size_P1 is the maximum three-dimensional diameter of the tumor, measured as the largest pairwise Euclidean distance between the voxels on the surface of the tumor volume; size_P2 and size_P3 are the tumor size along the directions orthogonal to the direction of maximum size (figure 2(b)).

Intensity features are first-order statistics that describe the distribution of the voxel intensities within the tumor on the CT image through commonly used basic metrics (Aerts et al 2014). Intensity features provide information related to the gray-level distribution of the image; however, they do not provide any information on the relative position of the various gray levels over the image. Therefore, we included textural features that either describe the patterns or the spatial distribution of voxel intensities, which were calculated from the gray level co-occurrence matrix (GLCM) (Haralick et al 1973). Texture matrix representation requires the voxel intensity values within the volume of interest to be discretized. In this work, voxel intensities were resampled into 64 equally spaced bins using a bin-width of 25 hounsfield units (HU). The detailed methodology of extracting intensity and texture features was previously described by Aerts et al (2014) (supplementary document). The calculated features of the 57 patients were normalized using the Z-scores method (Cheadle et al 2003).

| VHL         | PBRM1        | BAP1         |
|------------|--------------|--------------|
| Mutation   | Non-mutation | Mutation     | Non-mutation | Mutation     | Non-mutation |
| UTSW       | 26           | 7            | 19           | 14           | 5            | 28           |
| TCGA       | 10           | 14           | 3            | 21           | 2            | 22           |
| Total      | 36           | 21           | 22           | 35           | 7            | 50           |

Note: CT images from TCIA were acquired by the SIEMENS sensation 64/definition AS+ (Siemens Medical Solution, Malvern, PA), Philips Brilliance 64 (Philips Healthcare, Andover, MA), or GE LightSpeed VCT. CT image size was $512 \times 512$ with a pixel size of 0.7–1 mm and slice thickness was 1.25 or 5 mm.

2.2. MCMO predictive model

2.2.1. Evidential reasoning based classifier fusion

The evidential reasoning (ER) approach was used for fusing the individual classifier probability output (Yang and Xu 2002, 2013). Our study is a binary classification problem (mutation or non-mutation). Assuming there are $M$ classifiers, for a test sample, the output probability of each classifier is denoted by $P_i = \{P_{1i}, P_{2i}\}$, $i = 1, \cdots, M$, which satisfies:

$$P_{1i} + P_{2i} = 1. \quad (1)$$

Where $P_{1i}$ is output probability of gene mutation and $P_{2i}$ is output probability of non-mutation. Assume the relative weight of each classifier as $w = \{w_1, w_2, \cdots, w_M\}$, which satisfies the following constraint:
The final output probabilities $P_{fin}^j$, $j = 1, 2$ are obtained by classifier fusion through the ER approach (Yang and Xu 2002, 2013):

$$P_{fin}^j = ER \left( P_i^j, w_i \right), \quad i = 1, \cdots, M, \ j = 1, 2,$$

where ER represents the ER analytic algorithm (Wang et al 2006), which is calculated as:

$$P_{fin}^j = \frac{\mu \times \left[ \prod_{i=1}^{M} \left( w_i P_i^j + 1 - \omega_i \right) - \prod_{i=1}^{M} \left( 1 - w_i \right) \right]}{1 - \mu \times \left[ \prod_{i=1}^{M} \left( 1 - w_i \right) \right]}, \quad j = 1, 2,$$

where $\mu$ is calculated as:

$$\sum_{i=1}^{M} w_i = 1, \quad 0 \leq w_i \leq 1.$$
\[ \mu = \left[ \sum_{j=1}^{2} \prod_{i=1}^{M} \left( w_j p_j^i + 1 - w_j \right) - \prod_{i=1}^{M} (1 - w_i) \right]^{-1}. \]  

(5)

For a binary classification problem, if \( P_{1\text{fin}} > P_{2\text{fin}} \), the test sample belongs to class 1; if \( P_{1\text{fin}} < P_{2\text{fin}} \), the test sample belongs to class 2; if \( P_{1\text{fin}} = P_{2\text{fin}} \), the test sample belong to either class. In this study, we sought to predict either the presence or absence of gene mutation. Therefore, if \( P_{1\text{fin}} > P_{2\text{fin}} \), we considered the test sample had mutation; if \( P_{1\text{fin}} \leq P_{2\text{fin}} \), we considered the test sample had non-mutation.

2.2.2. Reliable outcome prediction based on output probability similarity

To obtain more reliable predictive results, reliable outcome prediction (RCP) is proposed and defined as maximize the similarity between predicted output probability and true label vector. For example, we assume two models that predict the VHL gene mutation with the label vector \([1, 0]\). Model A has prediction probabilities \((0.8, 0.2)\) (the probability of mutation is 0.8, the probability of non-mutation is 0.2, and the threshold is 0.5), and model B has prediction probabilities \((0.55, 0.45)\). As 0.8 is closer to 1 than 0.55, the result of model A is more reliable than that of model B. In other words, the similarity between the probability output of model A and the label vector is higher than those observed for model B, which means model A is more reliable than model B in this prediction.

In RCP, the aim is to maximize the similarity between predicted output probability and true label vector \(T\) while training the single classifier model and weights. For a training sample, its label vector is denoted by \( T = [T_1, T_2] \). \( T \) is a binary vector, \( T = [1, 0] \) (mutation) or \( T = [0, 1] \) (non-mutation). Assuming that the predictive model has \( q \) parameters denoted by \( R = \{ R_1, R_2, \cdots, R_q \} \), the objective function is expressed as:

\[ f = \max_{w, R} \sum_{k=1}^{K} \text{sim} \left( P^k, T^k \right), \]  

(6)

where \( K \) represents the number of training samples and \( \text{sim} \) is the similarity measure. Since the above problem can be considered as the similarity of probability distribution and the dice coefficient (Cha 2007) is effective for measuring similarity, it is used here:

\[ \text{sim} \left( P^k, T^k \right) = \frac{2 \sum_{j=1}^{3} p_j^k t_j^k}{\sum_{j=1}^{2} (p_j^k)^2 + \sum_{j=1}^{2} (t_j^k)^2}. \]  

(7)

Since a single objective may not be a good measure when the training dataset is imbalanced, we consider sensitivity and specificity simultaneously as a better solution as follows:
\[ f_{\text{sen}} = \frac{TP}{TP + FN}, \quad f_{\text{spe}} = \frac{TN}{TN + FP}, \]  

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. In our previous study, \( f_{\text{sen}} \) and \( f_{\text{spe}} \) were both considered as objective functions (Zhou et al. 2017).

However, \( f_{\text{sen}} \) and \( f_{\text{spe}} \) are label based measures, while we sought to maximize the similarity of probability output and the true label vector. Therefore, we defined the new similarity-based sensitivity and specificity denoted by \( f_{\text{sim}\_\text{sen}} \) and \( f_{\text{sim}\_\text{spe}} \), respectively. Assume that \( \{p_{t_1}^1, p_{t_2}^2, \cdots, p_{t_p}^p\} \) represents the probability output of true positives and the corresponding true label vector is \( \{T_{t_1}^1, T_{t_2}^2, \cdots, T_{t_p}^p\} \). The similarity of true positives \( TP_{\text{sim}} \) is defined as:

\[
TP_{\text{sim}} = \sum_{k=1}^{TP} \text{sim} \left( p_{t_k}^k, T_{t_k}^k \right) = \sum_{k=1}^{TP} \frac{2}{\sum_{j=1}^{2} \left( p_{t_j}^k \right)^2 + \sum_{j=1}^{2} \left( T_{t_j}^k \right)^2}. 
\]

In gene mutation prediction, \( j = 1 \) represents mutation, \( j = 2 \) represents non-mutation, and \( p_k \) is the mutation probability. Also, \( p_{t_1}^1 = p_k, p_{t_2}^2 = 1 - p_k, T_{t_1}^1 = 1, \) and \( T_{t_2}^2 = 0 \). Therefore, equation (9) can be simplified as:

\[
TP_{\text{sim}} = \sum_{k=1}^{TP} \frac{p_k^2}{p_k^2 + 1}. 
\]

Similarly, we define the similarity of true negatives \( TN_{\text{sim}} \), false positives \( FP_{\text{sim}} \), and false negatives \( FN_{\text{sim}} \):

\[
TN_{\text{sim}} = \sum_{k=1}^{TN} \frac{1 - p_k}{(p_k)^2 - p_k + 1}, \quad FP_{\text{sim}} = \sum_{k=1}^{FP} \frac{1 - p_k}{(p_k)^2 - p_k + 1}, \quad FN_{\text{sim}} = \sum_{k=1}^{FN} \frac{p_k}{(p_k)^2 - p_k + 1}. 
\]

Then, \( f_{\text{sim}\_\text{sen}} \) and \( f_{\text{sim}\_\text{spe}} \) are calculated as:

\[
f_{\text{sim}\_\text{sen}} = \frac{TP_{\text{sim}}}{TP_{\text{sim}} + FN_{\text{sim}}}, \quad f_{\text{sim}\_\text{spe}} = \frac{TN_{\text{sim}}}{TN_{\text{sim}} + FP_{\text{sim}}}. 
\]

Our aim was to maximize the two similarity-based objective functions simultaneously as:

\[
f_{\text{sim}} = \max_{w,R} (f_{\text{sim}\_\text{sen}}, f_{\text{sim}\_\text{spe}}). 
\]

Once training is finished, the Pareto-optimal solution set is generated, and the best model parameters and weights are selected based on the clinical needs. In the following subsection, we describe a new algorithm that was developed to solve the similarity-based multi-objective optimization problem.

### 2.3. Similarity-based multi-objective optimization (SMO) algorithm

Multi-objective evolutionary algorithms (MOEA) have demonstrated the superior performance for multi-objective optimization (Deb 2001). Based on MOEA, we have proposed an iterative multi-objective immune algorithm (IMIA), which adopts the traditional sensitivity and specificity as the optimized objective functions (Zhou et al. 2017). Based on IMIA, we propose a new SMO algorithm. The major difference between IMIA and SMO is that the reliability is measured through similarity between output probability and label vector during the training process. Additionally, weighting coefficients need to be optimized in SMO. For conciseness, we just gave a brief description of SMO and focused on the difference between SMO and IMIA. For a full and detailed algorithm, please refer to our previous paper (Zhou et al. 2017).

As IMIA, SMO consists of the seven steps: initialization, cloning, mutation, deletion, solution updating, termination and best solution selection. In initialization, model parameters \( R \) and weights \( w \) were both initialized. We generated the initial solution set \( D(t) = \{d_1, \cdots, d_H\} \) \( (t = 0) \) randomly, \( d_i(i = 1, \cdots, H) \) is a particular solution, \( H \) is the number of solutions, and \( t \) is the number of generation. In cloning step, there is a big difference. In SMO, new similarity-based proportional cloning operation was proposed, where the solution with higher similarity was reproduced multiple times. Specifically, the clonal time \( CLT \) for each solution is calculated as:
\[
\text{CLT}_i = n_t \times \frac{\text{sim}(d_i)}{\sum_{i=1}^{D} \text{sim}(d_i)},
\]

where \( n_t \) is the expected value of the clonal solution set and \( \lceil \cdot \rceil \) is the ceiling operator. The similarity measure for solution \( d_i \) denoted by \( \text{sim}(d_i) \) is calculated as:

\[
\text{sim}(d_i) = \sum_{k=1}^{K} \frac{2 \sum_{j=1}^{2} (P^k_j)^2 + 2 \sum_{j=1}^{2} (T^k_j)^2}{2^{2K}},
\]

where \( K \) is the number of training samples and \( T^k_j \) is the label vector.

The mutation and deletion in SMO are the same as those in IMIA. In this paragraph, ‘mutation’ refers to the operation performed on the cloned solution set, not the gene mutation. The mutation probability threshold \( MP \) is determined empirically and an operation probability \( RP \) is generated randomly. If \( RP > MP \), a mutation operation is performed in which a new solution \( d_i^m \) is generated randomly and replace the original solution \( d_i \). Then, a newly generated mutated solution set \( M(t) \) and solution set \( D(t) \) constitute the new solution set denoted by \( F(t) \). Same solutions in \( F(t) \) were removed and a new solution set \( DF(t) \) is generated.

In solution updating step of SMO, the aim is to select \( H \) solutions from \( DF(t) \) to maintain the population size. For each solution, we can obtain the similarities based on the probability outputs of all the training samples, according to equation (17). \( f_{sim\_sen} \) and \( f_{sim\_spe} \) can also be calculated according to equation (13). Using the MOEA (Deb 2001, Deb et al. 2002), we selected \( H \) solutions. Unlike most traditional MOEAs, the solution in \( DF(t) \) is sorted according to the similarity of each solution. Then, the new solution set \( UD(t) \) is generated.

When \( t \) reaches the maximal number of generations \( G_{max} \), the algorithm terminates. The best solution is selected from the Pareto-optimal solution set \( UD(G_{max}) \) according to clinical needs. We selected the best solution according to the similarity-based sensitivity, specificity, and AUC. First, the thresholds \( T_{sim\_sen} \) and \( T_{sim\_spe} \) are determined for similarity-based sensitivity and specificity according to clinical needs. Second, for each solution \( d_i \) in \( UD(G_{max}) \), we calculate its similarity-based sensitivity \( f_{sim\_sen} \) and specificity \( f_{sim\_spe} \). If \( f_{sim\_sen} > T_{sim\_sen} \) and \( f_{sim\_spe} > T_{sim\_spe} \), \( d_i \) is selected as a candidate solution. Third, we select the solution with the highest AUC from the candidate solutions as our final solution.

2.4. Training and testing procedure of the MCMO model

The training process mainly consists of three stages: feature calculation, feature selection, and predictive model construction. To achieve optimal performance for each classifier, we adopted our multi-objective feature selection method (Zhou et al. 2017). After selecting the features for each classifier, model parameters \( R_i \) and weights \( w_i \) \((i = 1, 2, \ldots, M) \) were trained. The workflow is illustrated in figure 3.

The testing process consists of three stages (figure 4). For a test sample, first, the features for each classifier are selected; second, each classifier outputs a probability \( P^i_j \) \((j = 1, 2; i = 1, 2, \ldots, M) \); third, the final mutation probability \( P^i_{fm} \) is obtained by combining all \( P^i_j \) and \( w_i \) using the ER approach. Then the label can be determined.

We used six different classifiers in the MCMO model, including support vector machine (SVM) (Keerthi and Lin 2003), logistic regression (LR) (Freedman 2009), discriminant analysis (DA) (Hastie and Tibshirani 1996), decision tree (DT) (Breiman 2001), K-nearest-neighbor (KNN) (Keller et al. 2012), and naive Bayesian (NB) (Goldszmidt 1997). Since SVM has two model parameters and other classifiers use default parameters, we train eight parameters \( \{ R_{SVM\_1}, R_{SVM\_2}, w_1, \ldots, w_6 \} \) for the predictive model.

3. Results

3.1. Experimental setup

In MCMO, the population number \( H \) and the maximal generation number \( G_{max} \) were both set to 100. In the clone operation, \( n_t \) was set to 200. In the mutation operation, the mutation probability \( MP \) was set to 0.9. The proposed MCMO predictive model was compared to three types of predictive model: (1) models with single classifiers; (2) models with different multi-objective optimization; (3) models with different fusion strategies. Because multi-objective optimization is better than the single-objective model (Zhou et al. 2017), we did not compare our proposed model to models with single-objective optimization. Area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, similarity-based sensitivity (Sim-sensitivity), and similarity-based specificity (Sim-specificity) were used for evaluating the model performance.
In our study, eight parameters in the predictive model need to be trained and our dataset has 57 cases. We adopted two-fold cross-validation in this work. Cross-validation is a widely used model validation technique which can test the model’s ability to predict new data that were not used in training, in order to flag problems like overfitting (Kohavi 1995). One round of cross-validation involves partitioning a sample of data into complementary subsets. In our work, two-fold cross-validation was used, where for each round, half cases (training set) were selected randomly for training and the other half cases (validation set) were used for validation, then reverse. To reduce variability caused by subset partition, ten rounds of two-fold cross-validation were performed for each model, and the validation results were averaged over the ten rounds to give an estimate of the model’s predictive performance (table 3). Prediction results of AUC, accuracy, sensitivity and specificity of training set were also listed in table 3. The prediction accuracy of training set is higher than that obtained from the validation set, which is considered as normal for a machine learning algorithm. On the other hand, the accuracies of the training set are not close to 1 and the predictive results on the validation set are acceptable. In the remaining of the paper, all the prediction results were from the validation set.

3.2. Selected features for the three genes

We obtained the classifier-specific feature subset and summarized the features selected by all (six) classifiers (table 4). Statistics using the unpaired $t$-test were also listed. The feature selected by all classifiers indicates that the feature is important to predict mutation. $P$-value smaller than 0.05 indicated significant differences between presence and absence of mutations. However, four selected features (Minimum, Contrast, Variance and Sum variance) for the $PBRM1$ gene and three selected features (Variance, Sum average and Sum variance) for the $BAP1$ gene had $P$-values greater than 0.05. Because we selected the optimal features according to multi-objective model, using the AUC as the figure of metric, these selected features did not necessarily have $P$-values smaller than 0.05. Example boxplots were plotted to show the potential of a single feature for differentiating between the presence and absence of mutations (figure 5).

Two intensity features including Mean and Kurtosis are the most frequently selected features in $VHL$ gene prediction. A histogram with a more elongated tail indicates smaller Kurtosis. A tumor with smaller Kurtosis is more likely to carry a $VHL$ mutation (figures 5(a) and 6). For the $PBRM1$ gene prediction, nine features from intensity and texture features, were most frequently selected. A boxplot of the Mean is illustrated in figure 5(b). Six features were selected as the most prominent contributors in $BAP1$. Four texture features were selected as follows: Homogeneity, Variance, Sum average, and Sum variance. A lower similarity in intensity between a voxel and
its neighbors led to higher Variance and Sum variance. A less uniform or more focal intensity distribution led to reduced Homogeneity. Therefore, larger Variance, and smaller Homogeneity were associated with the likelihood that a tumor carried a BAP1 mutation. Boxplot of Homogeneity is illustrated in figure 5(c).

It is noted that one single feature may not achieve accurate predictive results. For each classifier, a feature set is necessary. For KNN classifier, a feature set consisting of 12 features (Volume, Size_P3, Minimum, Maximum, Mean, Sum, Variance, Standard deviation, Kurtosis, Cluster shade, Energy, Inverse difference) was selected to predict VHL mutation; while for LR classifier, a feature set consisting of 19 features (Volume, Size_Z, Size_P2, Roundness, Surface area, Mean, Standard deviation, Skewness, Kurtosis, Contrast, Dissimilarity, Energy, Entropy, Homogeneity, Sum entropy, Information measure of correlation_1, Information measure of correlation_2, Inverse difference, Inverse difference normalized) was selected to predict VHL mutation.

### 3.3. Performance evaluation of MCMO versus single classifiers

MCMO yielded better AUC, accuracy, sensitivity, and specificity results than other single classifiers (table 5). The prediction accuracy of MCMO is 0.81, 0.78, and 0.90 for VHL, PBRM1, and BAP1 genes, respectively, with AUC $\geq 0.86$ sensitivity $\geq 0.75$ and specificity $\geq 0.80$. MCMO yielded better results than other single classifiers. KNN is the best single classifier for the VHL and PBRM1 genes, with specificities of 0.66 and 0.62, respectively. MCMO can achieve specificities of 0.86 and 0.80 for VHL and PBRM1 genes, respectively. SVM and DA achieved similar results for the BAP1 gene, which are better than other single classifiers, but sensitivities were only 0.57 and 0.63, respectively; the sensitivity obtained by MCMO was 0.87. Some single classifiers achieved higher sensitivities for the VHL gene, but the corresponding specificities were poor. MCMO achieved the highest AUC.

#### Table 3. Results of MCMO prediction models (training set versus validation set).

| Gene | Data set     | AUC       | Accuracy | Sensitivity | Specificity |
|------|--------------|-----------|----------|-------------|-------------|
|      | Training set | 0.96 ± 0.02 | 0.91 ± 0.01 | 0.93 ± 0.02 | 0.88 ± 0.02 |
|      | Validation set | 0.88 ± 0.01 | 0.81 ± 0.02 | 0.79 ± 0.04 | 0.86 ± 0.02 |
| VHL  | Training set | 0.95 ± 0.01 | 0.90 ± 0.03 | 0.87 ± 0.03 | 0.92 ± 0.02 |
|      | Validation set | 0.86 ± 0.02 | 0.78 ± 0.02 | 0.75 ± 0.02 | 0.80 ± 0.02 |
| PBRM1 | Training set | 0.97 ± 0.01 | 0.92 ± 0.01 | 0.9 ± 0.03  | 0.92 ± 0.01 |
|      | Validation set | 0.93 ± 0.02 | 0.90 ± 0.02 | 0.87 ± 0.02 | 0.90 ± 0.03 |
| BAP1  | Training set | 0.97 ± 0.01 | 0.92 ± 0.01 | 0.9 ± 0.03  | 0.92 ± 0.01 |
|      | Validation set | 0.93 ± 0.02 | 0.90 ± 0.02 | 0.87 ± 0.02 | 0.90 ± 0.03 |

#### Table 4. Selected features by all (six) classifiers for VHL, PBRM1 and BAP1 genes.

| Geometry (P-value) | Intensity (P-value) | Texture (P-value) |
|--------------------|---------------------|-------------------|
| VHL Mean (0.020)   | Kurtosis (0.002)    |                   |
| PBRM1 Minimum (0.088) | Mean (0.020)     | Maximum probability (0.030) |
| Mean (0.020)       | Median (0.041)     | Variance (0.162)  |
|                   | Skewness (0.042)   | Sum variance (0.168) |
| BAP1 Size_X (0.006) | Sum (0.048)       | Homogeneity (0.023) |
|                   |                     | Variance (0.171)  |
|                   |                     | Sum average (0.159) |
|                   |                     | Sum variance (0.168) |

![Figure 5. Boxplots for (a) kurtosis (VHL), (b) mean (PBRM1), and (c) homogeneity (BAP1).](image)
and accuracy with balanced sensitivities and specificities (difference < 0.1). Also, MCMO is a stable predictive model because its standard deviations are much smaller than those of the single classifiers.

### 3.4. Comparative study of objective functions

One group of the Pareto-optimal solution set and the selected final solution in SMO is shown in figure 7. As described in section 2.3, the best solution was selected according to the similarity-based sensitivity, specificity

| Gene   | Classifier | AUC     | Accuracy | Sensitivity | Specificity |
|--------|------------|---------|----------|-------------|-------------|
| VHL    | SVM        | 0.71 ± 0.06 | 0.69 ± 0.05 | 0.88 ± 0.07 | 0.37 ± 0.11 |
|        | NB         | 0.33 ± 0.14 | 0.67 ± 0.05 | 0.84 ± 0.10 | 0.36 ± 0.16 |
|        | LR         | 0.73 ± 0.04 | 0.71 ± 0.05 | 0.73 ± 0.05 | 0.67 ± 0.09 |
|        | KNN        | 0.80 ± 0.04 | 0.78 ± 0.03 | 0.84 ± 0.04 | 0.66 ± 0.08 |
|        | DT         | 0.67 ± 0.06 | 0.68 ± 0.07 | 0.71 ± 0.08 | 0.61 ± 0.11 |
|        | DA         | 0.71 ± 0.05 | 0.66 ± 0.05 | 0.68 ± 0.11 | 0.64 ± 0.10 |
|        | MCMO       | 0.88 ± 0.01 | 0.81 ± 0.02 | 0.79 ± 0.04 | 0.86 ± 0.02 |
| PBRM1  | SVM        | 0.68 ± 0.06 | 0.63 ± 0.06 | 0.52 ± 0.14 | 0.70 ± 0.06 |
|        | NB         | 0.41 ± 0.09 | 0.58 ± 0.05 | 0.52 ± 0.11 | 0.62 ± 0.08 |
|        | LR         | 0.62 ± 0.10 | 0.63 ± 0.09 | 0.52 ± 0.14 | 0.69 ± 0.09 |
|        | KNN        | 0.67 ± 0.10 | 0.64 ± 0.10 | 0.65 ± 0.16 | 0.62 ± 0.11 |
|        | DT         | 0.52 ± 0.08 | 0.54 ± 0.07 | 0.44 ± 0.10 | 0.60 ± 0.10 |
|        | DA         | 0.59 ± 0.07 | 0.59 ± 0.06 | 0.54 ± 0.10 | 0.62 ± 0.09 |
|        | MCMO       | 0.86 ± 0.02 | 0.78 ± 0.02 | 0.75 ± 0.02 | 0.80 ± 0.02 |
| BAP1   | SVM        | 0.80 ± 0.07 | 0.81 ± 0.03 | 0.57 ± 0.11 | 0.85 ± 0.04 |
|        | NB         | 0.74 ± 0.11 | 0.02 ± 0.02 | 0.24 ± 0.0 | 0.88 ± 0.02 |
|        | LR         | 0.69 ± 0.08 | 0.81 ± 0.03 | 0.54 ± 0.21 | 0.84 ± 0.04 |
|        | KNN        | 0.72 ± 0.06 | 0.80 ± 0.05 | 0.57 ± 0.10 | 0.83 ± 0.06 |
|        | DT         | 0.54 ± 0.09 | 0.81 ± 0.06 | 0.24 ± 0.18 | 0.89 ± 0.08 |
|        | DA         | 0.82 ± 0.08 | 0.80 ± 0.05 | 0.63 ± 0.20 | 0.82 ± 0.04 |
|        | MCMO       | 0.93 ± 0.02 | 0.90 ± 0.02 | 0.87 ± 0.02 | 0.90 ± 0.03 |
First, thresholds $T_{sim\_sen}$ and $T_{sim\_spe}$ were determined for similarity-based sensitivity and specificity based on clinical needs. In this study, the thresholds $T_{sim\_sen}$ and $T_{sim\_spe}$ are both 0.9. The selected candidate solutions were included within the red rectangle and the selected final solution (highest AUC) was marked in red.

We evaluated the performance of our MCMO by comparing it to the iterative multi-objective immune algorithm (IMIA), which adopts the traditional sensitivity and specificity as the optimized objective functions (Zhou et al 2017) (figure 8). The two methods were compared with the unpaired $t$-test at a significance level 0.05 (table 6). Results were similar based on AUC, accuracy, sensitivity and specificity ($P$-value $> 0.05$). For VHL and PBRM1 genes, SMO achieved a little higher AUCs. For all three genes, SMO achieved significantly higher similarity scores ($P$-value $\leq 0.01$), indicating that these results are more reliable. For the prediction result with higher AUC, the difference of $f_{sim\_sen}$ or $f_{sim\_spe}$ between SMO and IMIA is small. For example, for the BAPI gene, the difference of $f_{sim\_sen}$ and that of $f_{sim\_spe}$ are 0.03 and 0.02. However, for the prediction result with lower AUC, such as for the PBRM1 gene, the difference of $f_{sim\_sen}$ is 0.13 and that of $f_{sim\_spe}$ is 0.07.

Figure 7. Pareto-optimal solution set (green rhombus), the green rhombus within the red rectangle are candidate solutions which satisfy $f_{sim\_sen} > T_{sim\_sen}$ and $f_{sim\_spe} > T_{sim\_spe}$. $T_{sim\_sen} = T_{sim\_spe} = 0.9$. The best solution selected for SMO is indicated by the red rhombus. (a) VHL; (b) PBRM1; (c) BAP1.
Figure 8. Results of using different objective functions. (a) VHL; (b) PBRM1; (c) BAP1.

Table 6. Results of $P$-values compared between SMO and IMIA.

| Gene  | AUC   | Accuracy | Sensitivity | Specificity | Sim-sensitivity | Sim-specificity |
|-------|-------|----------|-------------|-------------|----------------|-----------------|
| VHL   | 0.116 | 0.449    | 0.226       | 0.206       | 0.003          | 0.0002          |
| PBRM1 | 0.412 | 0.690    | 0.355       | 0.355       | <0.0001        | <0.0001         |
| BAP1  | 0.574 | 0.536    | 0.330       | 0.472       | 0.01           | 0.007           |
3.5. Comparative study of fusing method

We used the ER approach (equation (4)) for fusing the output of different classifiers. The classic weighted fusion (WF) method is used for comparison, as:

$$P = \sum_{i=1}^{M} P_iw_i$$  \hspace{1cm} (18)

where $P_i$ is the individual classifier output probability and $w_i$ is the relative weight. SMO is used in both fusion strategies, and the comparative results are shown in figure 9. The two methods were compared with the unpaired $t$-test at a significance level 0.05 (table 7). For VHL and PBRM1 genes, the ER approach achieved higher AUCs ($P$-value < 0.05). Also, sim-sensitivity and sim-specificity in ER are higher than WF ($P$-value ≤ 0.02), which indicates that more reliable results can be obtained when using ER fusion.

4. Discussion

The study of the association between diagnostic imaging features and mutations is a first critical step in the radiogenomics of ccRCC (Kuo and Yamamoto 2011). While the genomic landscape of ccRCC is first characterized by the loss of VHL function, recent advances in cancer genome sequencing have identified additional, prognostically significant mutations. Two hypothesis-generating studies indicated the potential association between individual CT features and mutations of the VHL gene, and also mutations in the PBRM1, BAP1, SETD2 and KDM5C genes (Karlo et al 2014, Shinagare et al 2015). The CT features identified by the radiologists were primarily morphological (e.g. necrosis, ill or well defined margins and renal vein invasion). In contrast, all of our features were quantitative descriptors extracted from the contoured tumor image. The feature extraction was automated, eliminating subjectivity and improving reproducibility.

The most frequently selected features varied depending on the gene. The most frequently selected features of the VHL gene were intensity features (Mean and Kurtosis), which described the mean values and intensity distribution in tumor volume. Intensity and texture features were found to be important in the PBRM1 predictive model. The most selected features of PBRM1 consisted of five intensity features that measured the intensity distribution in tumor volume, and four texture features that measured the local differences within an image. Karlo et al (2014) reported that nodular, heterogeneous enhancement and visibility of intratumoral blood vessels in tumors were more common among ccRCCs with underlying VHL mutations. Also, investigators found an association between a well-defined tumor margin and the VHL mutation, and an association between solid ccRCC and mutations in VHL and PBRM1. However, Shinagare et al (2015) did not observe any imaging characteristics associated with PBRM1 and VHL mutations. Because both evaluations were subjective and their features were morphological, we were unable to directly compare our quantitative results with those morphological features.

The most frequently selected features of BAP1 consisted of one geometry feature, one intensity feature, and four texture features. Texture features were the most prominent contributors in the BAPI predictive model. In a study by Shinagare et al (2015), the BAPI mutation was found to be associated with ill-defined tumor margins and calcification. We did not study tumor margins because they are not a quantitative feature. However, the presence of calcification may be associated with selected features such as Homogeneity and Variance.

We proposed a MCMO radiogenomics model that predicts gene mutations in ccRCC. Multi-classifier models can fully use information extracted by different classifiers and potentially improve prediction accuracy. We used all six different classifiers without considering the performance of individual classifiers. In future work, we will test the performance of single classifiers, and remove those with lower performance for the multi-classifier model. SVM and NB results for the VHL gene were found to be poor (table 5), therefore eliminating them and using LR, KNN, DT and DA for fusion may improve prediction results.

Both similarity-based sensitivity and specificity were considered simultaneously as the objects that guided construction of the predictive model. For the first time, we propose reliable outcome prediction, which refers to maximizing the similarity of output probability and true label (probability is 1). Higher similarity means higher reliability. We designed similarity-based sensitivity and specificity as optimized objective functions, which differ from traditional ones. Moreover, an SMO algorithm was developed to train the model to increase accuracy and confidence of predictive results. Compared with our previous IMIA method, similarity was adopted as a non-dominated sorting criterion upon updating the solution set. Also, the solution with higher similarity was kept. These findings indicate that the prediction results are more reliable (higher $f_{sim\_sen}$ and $f_{sim\_spe}$) when using our model. As for the fusion strategy, the ER approach is better than the classic WF in terms of similarity-based optimization. SMO algorithm and ER approach both contribute to reliability increase.

Our study presents a number of limitations. First, the number of patients used in our study is relatively small. In our model, eight parameters (two parameters of SVM and six weights) need to be estimated and two-fold cross-validation were used. While two-fold cross-validation results showed that our model achieved satisfactory results, a larger dataset and multi-classifier fusion could help to reduce the potential risk of overfitting (Dietterich
In a future work, we can apply this MCMO model in different classification where dataset of larger size is available. Second, recent advances in genetics have led to the identification of several mutations associated with ccRCC, including those involving the \textit{VHL}, \textit{BAP1}, \textit{PBRM1}, \textit{SETD2}, \textit{MUC4} and \textit{KDM5C} genes. The genomic information of all six genes were available for the 24 patients in TCGA/TCIA data collection. However, genomic information of the three genes (\textit{SETD2}, \textit{MUC4} and \textit{KDM5C}) was not available for most of the 33 UTSW cases used in the present study. Thus, we only considered \textit{VHL}, \textit{PBRM1} and \textit{BAP1} genes in this work. Third, the feature stability was not addressed in the current study. Our patient data was acquired by different CT scanners at different institutions with different protocols, which resulted in differences in pixel size and slice thickness, while the differences of scanner and protocol have influence on feature calculations (Mackin \textit{et al} 2015). Additionally,

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure9.png}
\caption{Results of using different fusion strategies (ER versus WF). (a) \textit{VHL}; (b) \textit{PBRM1}; (c) \textit{BAP1}.}
\end{figure}
tumor delineation was conducted by one physician and reviewed by another physician in this study. This could also introduce inter-observer variability in tumor delineation. Standardization of image acquisition protocols, automatic segmentation or consensus contours from more physicians may further improve the performance of the model developed in this work.

5. Conclusion

We proposed a multi-classifier multi-objective (MCMO) radiogenomics model that predicts VHL, PBRM1, and BAP1 gene mutations in ccRCC using quantitative CT feature set. Using our feature selection strategy and model, we achieved a predictive AUC greater than 0.85 for all three genes. Compared to single classifiers, multi-classifiers fused through ER and trained by developed SMO algorithm can greatly improve prediction accuracy and reliability. In MCMO, the concept of reliable outcome prediction was first proposed and applied to the optimization procedure, generating more reliable results. The MCMO model should not only be applied to radiogenomics, but also to solving other outcome prediction problems in medicine.

Acknowledgment

This work was supported in part by the American Cancer Society ACS-IRG-02-196 (J Wang), the US National Institutes of Health P50CA196516 (J Brugarolas, P Kapur, I Pedrosa) and R01CA154475 (I Pedrosa), and the National Natural Science Foundation of China 61401349 (X Chen) and 61571359 (X Mou). The authors thank Dr Damiana Chiavolini for editing the manuscript.

ORCID iDs

Xi Chen https://orcid.org/0000-0003-0944-8437
Xuanqin Mou https://orcid.org/0000-0003-1381-5260

References

Aerts H J, Velazquez E R, Leijenaar R T, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B and Rietveld D 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach Nat. Commun. 5 4006
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Cancer Genome Atlas Research Network 2013 Comprehensive molecular characterization of clear cell renal cell carcinoma Nature 499 43–9
Carles J, Chiriveilla I, Climent M A, Gallardo E, Del Alba A G, Maroto I P, Mellado B and Del Muro F X G 2012 Evaluation of patients with metastatic renal cell carcinoma after failure of first-line treatment Cancer Metastasis Rev. 31 53–9
Cheddle C, Wawer M P, Freed W J and Becker K G 2003 Analysis of microarray data using Z score transformation J. Mol. Diagn. 5 73–81
Clark K V B et al 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository J. Digit. Imaging 26 1045–57
Dalgliesh G L, Purge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C and Latimer C 2010 Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes Nature 463 360–3
Deh K 2001 Multi-Objective Optimization Using Evolutionary Algorithms vol 16 (New York: Wiley) pp 200–86
Deh K, Pratap A, Agarwal S and Meyarivan T 2002 New method for objective multiobjective optimization: NSGA-II IEEE Trans. Evol. Comput. 6 182–97
Dietterich T G 2000 Ensemble methods in machine learning Int. Workshop on Multiple Classifier Systems (Berlin: Springer) pp 1–15
Duns G, van den Berg E, van Duivenbode I, Oisinga J, Hollema H, Hofstra R M and Kok K 2010 Histone methyltransferase gene SETDB2 is a novel tumor suppressor gene in clear cell renal cell carcinoma Cancer Res. 70 4287–91
Freedman D A 2009 Statistical Models: Theory and Practice (Cambridge: Cambridge University Press) pp 1–25
Gerlinger M, Horswell S, Larkin J, Rowan A J, Salm M P, Varela I, Fisher R, McGranahan N, Matthews N and Santos C R 2014 Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing Nat. Genet. 46 225–33
Gerlinger M, Rowan A J, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A and Tarpey P 2012 Intradatum heterogeneity and branched evolution revealed by multiregion sequencing New Engl. J. Med. 366 883–92
Gevaert O, Echegaray S, Khuong A, Hoang C D, Shrager J B, Jensen K C, Berry G J, Guo H H, Lau C and Plevritis S K 2017 Predictive radiogenomics modeling of EGFR mutation status in lung cancer Sci. Rep. 7 411674
Gnarra J R, Torki Y, Weng Y, Schmidt L, Wei M H, Li H, Latif F, Liu S, Chen F and Duh F M 1994 Mutations of the VHL tumour suppressor gene in renal carcinoma Nat. Genet. 7 85–90
Goldszmidt M 1997 Bayesian network classifiers Mach. Learn. 29 131–63

Table 7. Results of P-values compared between ER and WF.

| Gene | AUC | Accuracy | Sensitivity | Specificity | Sim-sensitivity | Sim-specificity |
|------|-----|----------|-------------|-------------|----------------|----------------|
| VHL  | 0.021 | 0.083 | 0.760 | 0.018 | 0.007 | <0.0001 |
| PBRM1| 0.0001 | 0.244 | 0.306 | 0.492 | 0.04 | 0.02 |
| BAP1 | 0.882 | 0.045 | 0.330 | 0.074 | 0.02 | 0.0009 |
