Bioinformatics, 37(19), 2021, 3106–3114
doi: 10.1093/bioinformatics/btab242
Advance Access Publication Date: 30 April 2021
Original Paper

Genome analysis

A novel systematic approach for cancer treatment prognosis and its applications in oropharyngeal cancer with microRNA biomarkers

Shenghua He[1], Chunfeng Lian[2], Wade Thorstad[3], Hiram Gay[3], Yujie Zhao[4], Su Ruan[5], Xiaowei Wang [3,*] and Hua Li [4,6,7,*]

[1]Department of Computer Science and Engineering, Washington University in Saint Louis, St. Louis, MO 63130, USA, [2]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049 Shaanxi, China, [3]Department of Radiation Oncology, Washington University in Saint Louis, St. Louis, MO 63130, USA, [4]Carle Cancer Center, Carle Foundation Hospital, Urbana, IL 61801, USA, [5]Laboratoire LITIS (EA 4108), Equipe Quantif, University of Rouen, 76183 Rouen, France, [6]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and [7]Cancer Center at Illinois, Urbana, IL 61801, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on August 8, 2020; revised on March 10, 2021; editorial decision on April 11, 2021; accepted on April 12, 2021

Abstract

Motivation: Predicting early in treatment whether a tumor is likely to respond to treatment is one of the most difficult yet important tasks in providing personalized cancer care. Most oropharyngeal squamous cell carcinoma (OPSCC) patients receive standard cancer therapy. However, the treatment outcomes vary significantly and are difficult to predict. Multiple studies indicate that microRNAs (miRNAs) are promising cancer biomarkers for the prognosis of oropharyngeal cancer. The reliable and efficient use of miRNAs for patient stratification and treatment outcome prognosis is still a very challenging task, mainly due to the relatively high dimensionality of miRNAs compared to the small number of observation sets; the redundancy, irrelevancy and uncertainty in the large amount of miRNAs; and the imbalanced observation patient samples.

Results: In this study, a new machine learning-based prognosis model was proposed to stratify subsets of OPSCC patients with low and high risks for treatment failure. The model cascaded a two-stage prognostic biomarker selection method and an evidential K-nearest neighbors classifier to address the challenges and improve the accuracy of patient stratification. The model has been evaluated on miRNA expression profiling of 150 oropharyngeal tumors by use of overall survival and disease-specific survival as the end points of disease treatment outcomes, respectively. The proposed method showed superior performance compared to other advanced machine-learning methods in terms of common performance quantification metrics. The proposed prognosis model can be employed as a supporting tool to identify patients who are likely to fail standard therapy and potentially benefit from alternative targeted treatments.

Availability and implementation: Code is available in https://github.com/shenghh2015/mRMR-BFT-outcome-prediction.:

Contact: huali19@illinois.edu or xwang317@uic.edu

1 Introduction

Head and neck cancer is the fifth most common cancer in the United States (Street, 2019), with an overall survival rate lower than 50%. Although the incidence of other sub-sites of head and neck cancer has decreased steadily in the past decades, the number of oropharyngeal squamous cell carcinoma (OPSCC) cases has increased significantly (Ernster *et al.*, 2007; Gao *et al.*, 2013). Retrospective studies conducted by the International head and neck Cancer Epidemiology Consortium (INHANCE) have demonstrated that clinical biomarkers have prognostic value in helping stratify OPSCC patients into groups with varied risks of death or disease progression (Heck *et al.*, 2010; Winn, 2015). Human papillomavirus (HPV) is a known driving oncogenic factor in oropharyngeal cancer, as well as a significant prognostic biomarker for patient survival (Gillison *et al.*, 2008; Marur and Burtness, 2014). However, HPV-positive

oropharyngeal cancer patients have similar rates of metastatic spread to HPV-negative patients. The same is true for patient groups stratified with other clinical biomarkers (e.g. sex, age, tumor TNM stage and tumor size). There is an urgent need to determine oropharyngeal cancer's distinctive characteristics for patient stratification.

MicroRNAs (miRNAs) are a family of small non-coding RNA molecules that collectively controls the expression of thousands of protein-coding genes (Ambros, 2004; Chen *et al.*, 2018). Multiple studies indicate that miRNAs are promising biomarkers and play critical regulatory roles in oropharyngeal and other human cancers (Gao *et al.*, 2013; Miller *et al.*, 2015; Satapathy *et al.*, 2017). Among all human miRNAs, 533 are expressed in oropharyngeal tumors or normal oropharynx, as revealed by analyzing the Cancer Genome Atlas data (http://cancergenome.nih.gov/). Most reported miRNA studies focused primarily on the early diagnosis of head and neck cancer, but not the disease treatment outcome prognosis or survival analysis. The survival analysis is either to directly predict the risk of a patient/population, or to address the simpler binary classification problem (survived—not survived). The reliable and efficient usage of miRNAs for oropharyngeal cancer patient stratification with low and high risks of treatment failure remains a challenging problem, mainly due to the challenges described below. First, uncertainty of the profiled miRNAs with the corresponding outcome labels exist due to the heterogeneity of tumor tissues. Second, not all profiled miRNAs are useful and some of them might even mislead the patient stratification. Redundancy among the extracted miRNA biomarkers and irrelevancy of the miRNAs to outcomes exist. Third, imbalanced (skewed) dataset due to different treatment outcome rates can result in higher false positive rates on the patient cases with outcomes in the minor class. Fourth, relatively small training samples compared to the high-dimensional miRNA feature space may result in a high risk of over-fitting and decrease the prognosis performance on unseen patient data. Sparse and robust prognostic miRNAs are desired to stratify OPSCC patients for targeted treatment.

Prognostic miRNA biomarker identification can be considered as a problem of feature selection that needs solutions to address the above-mentioned challenges. Numerous feature selection methods have been proposed in the past decades (Cheng *et al.*, 2011; Kira and Rendell, 1992a; Kwak and Choi, 1999; Lin and Jeon, 2006; Sun, 2007). Some reported methods aimed to select informative features by considering feature-label relevance. For examples, Kira and Rendell (1992a) and Sun (2007) proposed RELIEF (RELevance In Estimating Features) and I-RELIEF (iterative RELIEF) algorithms to weight the relevance of features with class labels in terms of Euclidean distance for feature selection. Wang *et al.* (2012) employed feature-label correlation coefficients to select features, while Gao *et al.* (2013) utilized a cox proportional hazards model to select outcome-relevant miRNAs.

Differently, other methods have been proposed to select features by considering both feature-label relevance feature-feature redundancy. For examples, Eid *et al.* (2013) calculated Pearson correlation coefficients between features and labels, which were employed to select features with high feature-label relevance and low feature-feature redundancy based on a sequential-searching strategy. Hall (2000) have proposed a correlation-based method to rank the redundancy of feature subsets for feature subset selection instead of individual feature selection. Peng *et al.* (2005) have designed an iterative minimal-redundancy-maximal-relevance (mRMR) evaluation strategy, which employed mutual information to characterize the relationships between features and labels for feature selection. The statistical correlation measurements investigate variables' non-independence with their products. The features selected based on these measurements might still yield stocastically dependent. Instead of considering the linear covariance, mutual information evaluates variables' non-independence with their joint probability distributions, which provides more thoughtful evaluation of variables' dependence. Therefore, mutual information between features and labels are considered as a powerful tool to select features with low feature-feature redundancy and high feature-label relevance.

The above-mentioned methods, which exploited intrinsic feature-label relevance and feature-feature redundancy for feature selection, are suitable for a filtering (or embedding feature selection) situation without specifying following classifiers. However, only considering the feature-feature redundancy and feature-label relevance might select redundant and label-irrelevance features for specific classifiers and might increase the potential over-fitting risk on unseen data and decrease the classification performance. The majority of reported methods aimed to select informative features which optimize a pre-determined classifiers (Kennedy, 2006; Lian *et al.*, 2015; Mi *et al.*, 2015; Tang *et al.*, 2014). The informativeness of the selected features is also evaluated by the performance of the classifiers. Commonly, these methods select features and optimize the performance of the classifier simultaneously through the minimization of loss functions with a set of training dataset. When the dimension of the feature space is high and the size of training dataset is relative small (Loughrey and Cunningham, 2004), using all available features to optimize a loss function increases the computational burden and yields sub-optimization due to the feature-feature redundancy and feature-label irrelevance described above. Sparsity learning methods (Lian *et al.*, 2015; Tan *et al.*, 2010) have been employed to utilize prior knowledge for feature selection. These methods might still select irrelevant and redundant features without considering intrinsic properties of feature-label relevance and feature-feature redundancy.

Other reported methods consider both feature intrinsic properties and classification performance (El Akadi *et al.*, 2011; Ge *et al.*, 2016; Lian *et al.*, 2016b; Wen *et al.*, 2019). For example, Random Forests or Random Survival Forests (RSF) (Breiman *et al.*, 1984; Lin and Jeon, 2006) are state-of-the-art machine-learning methods which are known for handling large dimensionality datasets and performing the feature selection and survival prediction. However, RSF methods select informative features individually instead of determining the subset of features simultaneously. In a two-stage feature selection method proposed by Ge *et al.* (2016), a subset of informative features was first determined through the evaluation of feature-label relevance and feature-feature redundancy with a metric of maximal information coefficient (MIC). The feature subset was then refined by optimizing a pre-defined k-nearest neighbors (K-NN) classifier based on a best-first search strategy (Pearl, 1984). Lian *et al.* (2015) employed RELIEF method (Kira and Rendell, 1992a) to pre-select a feature subset based on feature-label relevance, and then refined the feature subset by optimizing an evidential K-NN classifier with Belief Function Theory (BFT) (Denoeux, 2008). RELIEF method was employed to select informative features individually based on feature-label relevance. However, redundant features might still be selected, which can affect the final classification performance. To the best of our knowledge, no reported methods have been proposed to address the above-mentioned four challenges.

In this study, a prognosis model was proposed to address the above described challenges to select informative feature subset and optimize a binary classifier for patient stratification, which was motivated by the previous work (Lian *et al.*, 2015; Wu *et al.*, 2019). The proposed model was employed to reliably stratify subsets of 150 OPSCC patients with low and high risks of treatment failure, and the performance was compared to other state-of-the-art methods. The potential applications of the proposed method was discussed as well.

2 Materials and methods

The proposed prognosis model, shown in Figure 1, included three steps of: (i) mRMR-based feature pre-selection, (ii) BFT-based feature refinement and (iii) evidential K-nearest neighbors (EK-NN) classifier. Given a training dataset including tumor samples with profiled miRNAs and treatment outcomes, the mRMR-based method first selected a subset of miRNAs that are most relevant to outcome labels and yield less redundancy with each other. The belief function theory (BFT)-based feature refinement method refined the pre-selected miRNAs to a highly sparse feature subset that addresses the issues of small and imbalance training datasets and class label uncertainty through the optimization of the pre-defined classifier. The EK-NN classifier was employed as the pre-defined classifier for

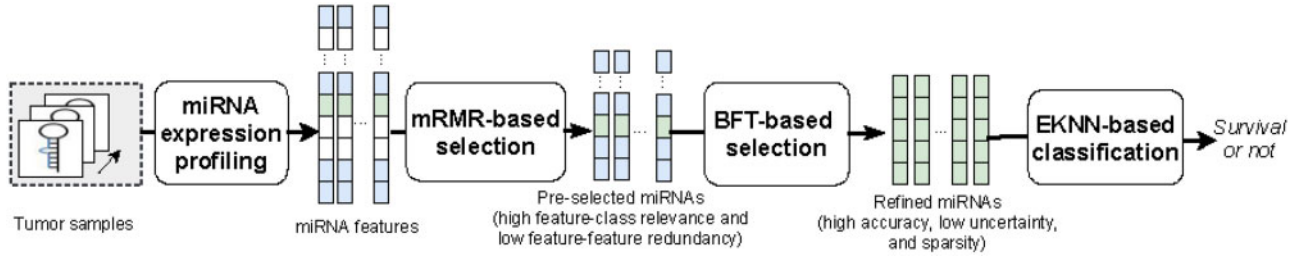

Fig. 1. Framework of the proposed prognosis model. The squares in each step represent the miRNA features. The white biomarkers represent those non-selected features. The mRMR-based feature selection method first selects those blue and green biomarkers which yield high feature-class relevance and low feature-feature redundancy. The BFT-based feature refinement method further determines those sparse biomarkers (green) which yield low feature-label uncertainty and high classification accuracy

feature refinement and was trained together with the feature refinement. The refined miRNA feature subset and the trained EK-NN classifier were employed to stratify unseen patients into groups of low or high-risk of treatment failures.

2.1 Dateset preparation: miRNA expression profiling

One hundred fifty oropharyngeal squamous cell carcinoma (OPSCC) patient cases were employed to demonstrate the model performance. The patient cases have been collected based on an Institutional Review Board (IRB) protocol approved by the Human Research Protection Office of the Washington University School of Medicine in St. Louis. All the patient cases have been treated with radiation therapy. Half of them have also received surgery and/or chemotherapy. The miRNA expression profiling has been performed on these FFPE tumor samples using our established real-time RT-PCR method (Wang, 2009) to determine 96 cancer-related miRNAs based on the dysregulation of these miRNAs in various human cancers. Sections from each tumor sample were stained with hematoxylin and eosin (H&E) and reviewed independently by two study pathologists at Washington University to confirm diagnoses. The expression levels of individual miRNAs profiled from each tumor sample were normalized for the study. More details of the profiling procedure was explained by Gao *et al.* (2013).

The age of patient cases at the time of diagnosis ranged from 32 to 87 with an average of 56.5 years. The patients are predominantly Whites (86%) and the rest patients are Africa-American (12%) and Native Americans and Asians (2%). All the patients were treated with radiotherapy (definitive or post-operative). Overall survival (OS) and disease-specific survival (DSS) were considered as two different types of treatment outcomes. OS period was defined as the time in between the date of treatment received and the date of death, which ranges from 67-4268 days. DSS is a net survival measure representing cancer survival in the absence of other causes of death, which estimates the probability of surviving using the definition of specific cause of death. DSS period was defined as the time between the date of treatment received and the date the patient survives without any symptoms of the OPSCC, which ranges from 1 to 4268 days. When stratifying the outcomes based on OS, 99 cases have the labels of OS and the rest 51 cases have the labels of non-OS. When stratifying the outcomes based on DSS, 96 cases have the labels of DSS and the rest 54 cases have the labels of non-DSS. Without loss of generality, the label of OS can be represented as ω_1 and the label of non-OS can be represented as ω_2 when stratifying patient cases based on OS. The label of DSS can be represented as ω_1 and the label of non-DSS can be represented as ω_2 when stratifying patient cases based on DSS. The labels ω_1 and ω_2 will be used in the following sections for method description. All patient cases were de-identified prior to analysis.

2.2 mRMR-based miRNA biomarker pre-selection

An iterative minimal-redundancy-maximal-relevance (mRMR) evaluation strategy (Peng *et al.*, 2005) was employed to pre-select profiled miRNAs based on the feature-label relevance and feature-feature redundancy. Let $X = \{x_1, x_2, \ldots, x_K\}$ represent the set of K profiled miRNAs, $\Omega = \{\omega_1, \omega_2\}$ represent the set of outcome labels related to X, $r(x_{k_1}, x_{k_2})$ is defined as the similarity of any two features x_{k_1} and x_{k_2}, $k_1, k_2 \in \{1, 2, \ldots, K\}$, and $v(x)$ is defined as the relevance of feature x to the outcome labels, $x \in X$. Both $r(x_{k_1}, x_{k_2})$ and $v(x)$ are defined as the discrete mutual information (Ding and Peng, 2005; Ross, 2014; Steuer *et al.*, 2002):

$$r(x_{k_1}, x_{k_2}) = \sum_{C_{k_1}} \sum_{C_{k_2}} p(x_{k_1}, x_{k_2}) \log \frac{p(x_{k_1}, x_{k_2})}{p(x_{k_1}) p(x_{k_2})}, \tag{1}$$

$$v(x) = \sum_{C_x} \sum_{C_\omega} p(x, \omega) \log \frac{p(x, \omega)}{p(x) p(\omega)}, \tag{2}$$

where C_{k_1}, C_{k_2}, C_x and C_ω represent all possible states that a measurement performed on x_{k_1}, x_{k_2}, x and ω ($\omega \in \Omega$), respectively. Here, $p(x_{k_1}, x_{k_2})$ represents the joint probabilistic distribution of x_{k_1} and x_{k_2}, while $p(x, \omega)$ the joint probabilistic distribution of x and ω, and $p(x)$ and $p(\omega)$ the marginal probabilistic distributions of x and ω, respectively. Given X, $p(x_{k_1}, x_{k_2})$, $p(x, \omega)$, C_{k_1}, C_{k_2}, C_x and C_ω are estimated by use of curve fitting (Ding and Peng, 2005; Ross, 2014). A subset of features $S \subseteq X$ is selected by minimizing a loss function $L_m(S)$, which is defined as:

$$L_m(S) = R_d(S) - R_v(S), \tag{3}$$

where $R_d(S)$ is defined as the mean of redundancies between any two features in S, and $R_v(S)$ is defined as the mean of relevance between any feature x in S and outcome label variable ω:

$$R_d(S) = \frac{1}{|S|^2} \sum_{x_{k_1}, x_{k_2} \in S} r(x_{k_1}, x_{k_2}), \tag{4}$$

$$R_v(S) = \frac{1}{|S|} \sum_{x \in S} v(x). \tag{5}$$

Given a training dataset, a subset S^* yielding high feature-label relevance and low feature-feature redundancy is determined by minimizing $L_m(S)$ in Equation 3 with an incremental searching strategy shown in Algorithm 1.

2.3 BFT-based miRNA biomarker refinement

The goal of feature refinement is to further select a feature subset S^{**} from the S^* determined by mRMR method described above. This refinement process relates to the performance of a classifier. The feature refinement process was designed based on BFT theory (Dempster, 2008; Shafer, 1976; Wu *et al.*, 2019), considering its ability of reasoning with uncertain and imprecise information by aggregating partial and uncertain evidences. BFT is a generalization of both probability theory and set-membership approaches, and closely relates to imprecise probability (Walley, 2000) and random sets (Nguyen, 2006). The traditional Bayesian classification considers the probability of a sample's label belonging to each class, but BFT-based feature refinement method considers not only the probability of a sample belonging to each single label but also belonging to the subsets of all labels, which can deal with class label imprecision and uncertainty (the challenges described in this study).

In this study, four possibilities of a sample's class label were considered and defined as $A \subseteq \{\omega_1, \omega_2, \Omega, \varnothing\}$, in which $\varnothing$ represents an empty label set. In this way, the uncertainty of a sample' label is handled more precisely than only separating the possibilities of a sample's label belonging to either ω_1 or ω_2. Let ω denote the label of a sample X, the evidence regarding the actual value of ω can be represented by a mass function m on Ω, which was defined from the power set 2^{Ω} to the interval $[0, 1]$:

$$\sum_{A \in \Omega} m(A) = 1, \tag{6}$$

where $m(A)$ denotes *a degree of belief* attached to the hypothesis that '$\omega \in A$'. The mass function induced by a sample X_i, which supports the assumption that another sample X_j has the same class label of X_i, is defined as:

$$\begin{cases} m_{ij}(\{\omega_c\}) & = \alpha e^{-\gamma_c d_{ij}^2} \\ m_{ij}(\Omega) & = 1 - \alpha e^{-\gamma_c d_{ij}^2} \end{cases}, \tag{7}$$

where $c = \{1, 2\}$, α and γ_c are the weight factors, and $d_{i,j}$ represents the distance between X_i and X_j:

$$d_{i,j}^2 = \sum_{v=1}^{V} \lambda_v d_{ij,v}^2, \tag{8}$$

where $d_{ij,v} = |x_{i,v} - x_{j,v}|$ represents the Euclidean distance between the vth feature of X_i and that of X_j, $1 \leq v \leq V$, and λ_v represents an element in Λ. Large $d_{ij,v}$ represents negligible information provided by the vth feature of X_i to X_j. Given a set of N samples X_n, $1 \leq n \leq N$, which can be separated into two groups of samples, Θ_1 and Θ_2, with different labels ω_1 and ω_2, respectively. The mass function $m_i^{\Theta_c}$ induced by the set of N samples, which supports the assumption that the sample X_i has the same class label of Θ_c, is defined as:

Algorithm 1: mRMR-based miRNA feature pre-selection strategy

Input :

- A set of N training data, in which $\{X_n, \omega_n\}$ represents the n^{th} training sample, $X_n \in \mathbb{R}^K$ represents the feature vector of the n^{th} training sample, K represents the size of features, and $\omega_n \in \Omega$ is its associated outcome label of X_n, $n \in \{1, 2, ..., N\}$, and $\Omega = \{\omega_1, \omega_2\}$;
- A pre-defined empty feature set S^* with the size $V = 0$.

1. Employ the N training data to estimate the probability distributions $p(x)$, $p(\omega)$, $p(x_{k_1}, x_{k_2})$, and $p(x, \omega)$ [40];
2. For $\forall x \in X$, compute the feature-label relevancy $v(x)$ based on Eqn. 2;
3. Select the feature x^* that yields the maximum relevancy $v(x)$, $x^* = \arg\max_{x \in X} v(x)$;
4. Initialize $S^* = \{x^*\}$, $V = V + 1$;
5. Calculate $L_m(S^*) = \frac{1}{|S^*|} \sum_{x_s \in S^*} r(x, x_s) - \frac{1}{|S|} \sum_{x \in S} v(x)$;

while $V \leq K$ **do**
  Update the miRNA feature set $\bar{X} = X - S^*$;
  for $\forall x \in \bar{X}$ **do**
    Calculate $r(x, x_s)$ for $\forall x_s \in S^*$ based on Eqn. 1;
    Calculate $\Delta L_m(x) = \frac{1}{|S^*|} \sum_{x_s \in S^*} r(x, x_s) - v(x)$;
  $x^* = \arg\min_{x \in \bar{X}} (\Delta L_m(x))$;
  $S^{*\prime} = S^* \cup \{x^*\}$;
  $L_m(S^{*\prime}) = R_d(S^{*\prime}) - R_v(S^{*\prime})$;
  if $L_m(S^{*\prime}) < L_m(S^*)$ then $S^* = S^{*\prime}$;
  $V = V + 1$;

Output: The selected feature subset set S^*.

$$\begin{cases} m_i^{\Theta_c}(\{\omega_c\}) & = 1 - \prod_{X_n \in \Theta_c}^{n=1,...,N} (1 - e^{-\gamma_c d_{i,n}^2}) \\ m_i^{\Theta_c}(\Omega) & = \prod_{X_n \in \Theta_c}^{n=1,...,N} (1 - e^{-\gamma_c d_{i,n}^2}) \end{cases}, \tag{9}$$

where $c = \{1, 2\}$, $m_i^{\Theta_c}(\Omega) = 1$ when Θ_c is empty. A global mass function M_i regarding the class membership of X_i can be calculated as:

$$\begin{cases} M_i(\{\omega_1\}) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\Omega) \\ M_i(\{\omega_2\}) & = m_i^{\Theta_2}(\{\omega_2\}) \cdot m_i^{\Theta_1}(\Omega) \\ M_i(\Omega) & = m_i^{\Theta_1}(\Omega) \cdot m_i^{\Theta_2}(\Omega) \\ M_i(\varnothing) & = m_i^{\Theta_1}(\{\omega_1\}) \cdot m_i^{\Theta_2}(\{\omega_2\}) \end{cases}. \tag{10}$$

Defining a binary vector Λ with the size of the number of elements in S^*, the goal of feature refinement process is to determine the value of an element λ in Λ to be either 0 or 1. In the determined Λ, 1 represents the corresponding feature in S^* was selected, while 0 represents the feature was not selected. The refined feature subset S^{**} is those features corresponding to $\lambda = 1$. Given the definition of mass function and global mass function, each λ in Λ can be determined through the minimization of a loss function $L_b(\Lambda)$:

$$L_b(\Lambda) = L_{\text{err}}(\Lambda) + L_{\text{uncertain}}(\Lambda) + \beta ||\Lambda||_0. \tag{11}$$

The first item $L_{\text{err}}(\Lambda)$ in $L_b(\Lambda)$ measures the mean squared error between the predicted probability scores and the outcome labels in the feature subspace determined by Λ. Here $L_{\text{err}}(\Lambda)$ was defined as:

$$L_{\text{err}}(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{2} (M_n(\{\omega_c\}) - y_{n,\omega_c})^2, \tag{12}$$

where $M_n(\{\omega_c\})$ represented a global mass function that quantifies the level of evidence that sample X_n has class label of ω_c, $y_{n,\omega_1} = 1$ and $y_{n,\omega_2} = 0$ when the label of X_n was ω_1, and $y_{n,\omega_1} = 0$, and $y_{n,\omega_2} = 1$ when the label of X_n was ω_2.

The second item $L_{\text{uncertain}}(\Lambda)$ in $L_b(\Lambda)$ was defined as:

$$L_{\text{uncertain}}(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} (M_n(\Omega)^2 + M_n(\varnothing)^2), \tag{13}$$

where $M_n(\varnothing)$ defined in Equation (10) measures the conflict in the neighborhood of X_n, and $M_n(\Omega)$ measures the imprecision regarding the class membership of X_n (Wu *et al.*, 2019). The item L_{err} and item $L_{\text{uncertain}}$ work together to select the features which can correctly estimate samples' labels and penalize a feature subset that results in conflict and imprecise evidence. The features are selected by considering both the uncertainties of features and class labels and other issues through the training process.

Last term $\beta ||\Lambda||_0$ is a sparsity constraint, in which $||\Lambda||_0 = \sum_{v=1}^{V} \lambda_v$ represents the number of non-zero entries in Λ and β is a scalar that controls the strength of the sparsity penalty. This term forces the selected miRNA feature subset to be sparse in order to decrease the over-fitting risk on unseen data, and lead to high classification accuracy and small overlaps between different classes.

2.4 EK-NN classifier for patient stratification

The EK-NN classifier (Denoeux, 1995; Denœux and Kanjanatarakul, 2016; Lian *et al.*, 2016a; Liu *et al.*, 2017; 2018) was employed as the pre-defined classifier for the feature refinement and the final classifier for stratifying unseen data as well. The original voting K-NN (Dudani, 1976) assigns a sample into the class represented by its majority nearest neighbors in the training set without concerning the dissimilarity (distance) between the sample and its neighbors. To endow the K-NN method with the capability to consider the sample dissimilarity to better handle the uncertain information, the EK-NN rule provides a global treatment of partial knowledge regarding the class membership of training patterns. Ambiguity and distance reject options are also taken into account based on the concepts of lower and upper expected losses (Quost

et al., 2011). The EK-NN method has outperformed other traditional K-NN methods in many situations when using the same information (Zouhal and Denoeux, 1998).

The parameters α and γ_c in Equation (7) were determined along with the minimization of the loss function defined in Equation (11). The loss function $L_b(\Lambda)$ was minimized by use of an integer genetic algorithm (Damousis *et al.*, 2004). The determined parameters α and γ_c were then employed to calculate the mess-functions in Equation (7) which were employed to measure the differences (mass-function based distances) of a given sample to its neighbors and stratify it. In addition, to overcome the challenge of feature selection with imbalance data, an adaptive synthetic sampling (ADASYN) (He and Garcia, 2008) was employed to rebalance data by generating synthetic minority class samples according to their distribution. The key idea of ADASYN is to create synthetic samples according to the distribution of the minority class samples, where more samples are generated for the minority class samples that have higher difficulty in learning. The level of difficulty in learning for each minority samples is measured by the ratio of the majority class samples in each minor class sample's k-nearest-neighborhood. In this study, five neighbors are considered for the minor class data rebalance according to the processed data. ADASYN outputs a balanced training dataset via the procedure described in the literature (He and Garcia, 2008; Lian *et al.*, 2015).

2.5 Prognosis model training and validation

In this study, overall survival (OS) and disease-specific survival (DSS) were employed as the end points of treatment outcome, respectively. The proposed method was trained and tested separately to stratify low-risk and high-risk patients based on either OS or DSS. Of all 150 samples, 101 samples were randomly selected as the training set while the rest 49 samples were considered as the testing set based on OS labels. The training dataset included 40 OS positive and 61 negative cases, and the testing dataset included 11 positive and 38 negative cases, respectively. Positive cases represent survival and negative represents not. The same separation strategy was also employed to determine training and testing datasets based on the DSS label, of which 41 positive and 60 negative cases were included in the training datasets and 13 positive and 36 negative cases were included in the testing datasets. Five-fold cross validation was employed to train and validate the proposed method. The training data was used alone without involving any testing data. The hyperparameters, α, β and the number of neigbors in EK-NN, were optimized individually through the line search. The model that achieved the lowest validation loss, which is the average loss over the folds, was employed for the final evaluation on the testing data. In addition, the searching range of the parameters was summarized in Table 1. The range is determined considering that (i) β should be much less than 1 for reasonable sparsity of features, (iii) α should be close to 1 considering small uncertainty of features and labels and (iii) the numbers of neighbors is chosen from the commonly setting of that in K-NN methods.

The prediction accuracy, F1 score, the area under the receiver operating characteristic curve (AUC), and Kalpan Meier survival curves (Bland and Altman, 1998) were employed to evaluate the performance of the proposed method. The prediction accuracy is defined as: Accuracy $= \frac{T_p+T_n}{T_p+T_n+F_n+F_n}$, and the F1 score is defined as: F1 $= \frac{2T_p}{2T_p+F_p+F_n}$. Here T_p is the number of positive cases that are correctly predicted as positive ones, T_n is the number of negative cases that are correctly predicted as negative ones, F_p is the number of the cases that are predicted to be positive but in fact are negative ones, and F_n is the number of the cases that are predicted to be negative but in fact are positive ones. The F1 score is an effective metric for evaluating prediction performance when having imbalanced testing dataset. The ROC curve and the area under the ROC curve (AUC) were also employed to visualize the stratification performance. The AUC was calculated based on the ROC curves fitted by use of the tools developed by Metz (1999). The Kaplan Meier curves were plotted by use of Kaplan–Meier analysis methods (Bland and Altman, 1998) and the open-source software (Creed *et al.*, 2020).

Table 1. The parameters to be determined through the training process

Parameters	Range of settings
β in Equation 11	determined from $\{0.001, 0.01, 0.02, 0.05, 0.07, 0.1\}$
No. of the neighbors in EK-NN	determined from $\{5, 7, 9, 11\}$
α	determined from $\{0.8, 0.85, 0.9, 0.95\}$
γ_c	$= \frac{1}{d_c^2}$ and determined through loss function minimization, where d_c^2 represents the mean distance between any two samples with the same ω_c

3 Results

Figure 2 showed the performance of the proposed method on stratifying patient cases based on OS and DSS as the endpoints of the outcome, respectively. It can be observed that the proposed method achieved high performance. The proposed model has been further evaluated through the comparison with other state-of-the-arts methods. The results were shown in Figures 3 and 4. The first compared method is the cox proportional hazards regression (CoxReg) analyses-based method (Cox, 1972; Gao *et al.*, 2013), which is one of the most popular regression techniques for survival analysis. The BFT-based evidential forward feature selection (EFS-BFT) method (Lian *et al.*, 2015), and the RELIEF-based BFT (RELIEF-BFT) method (Lian *et al.*, 2015) were also compared to the proposed mRMR-BFT method. In the EFS-BFT method, a BFT-based feature selection method was directly applied to all profiled miRNA features to select a sparse subset. In the RELIEF-BFT method, the RELIEF algorithm (Kira and Rendell, 1992b) was employed to pre-select informative features based on feature-label relevance, and then BFT-based feature refinement was applied to determine the final sparse subset of miRNA features. In the mRMR-BFT method, the feature set selected by the mRMR method was refined to a sparse feature set by use of the BFT method. The EK-NN classifier was employed as the predefined classifier to train all these compared methods excluding the CoxReg method. The EK-NN classifier, which was trained by use of all profiled features directly, was also compared to demonstrate the significance of feature selection process. The same training and testing data separation were employed for other compared methods for fair comparison. The parameters in all methods were optimized separately to achieve the best performance of each. The comparison results show that the proposed mRMR-BFT method achieved higher performance in terms of the metrics of prediction accuracy, AUC and F1 score.

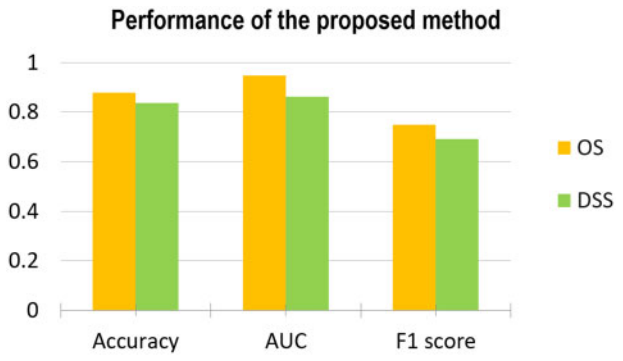

Fig. 2. Performance of the proposed method evaluated by use the matrices of Accuracy, AUC and F1 score and based on the outcome labels of OS (**a**) and DSS (**b**), respectively

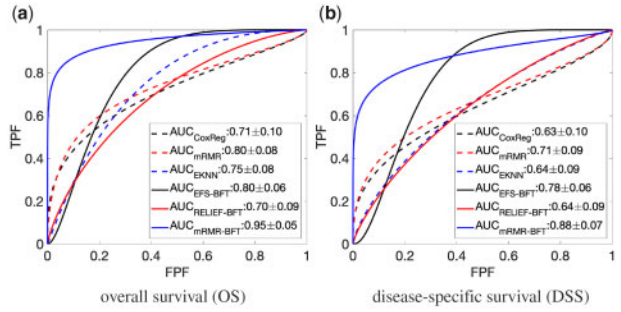

Fig. 3. Performance comparison of the proposed method and five other methods. The performance was evaluated by use of ROC curves and AUC values and considering the overall survival OS (**a**) and disease-specific survival DSS (**b**) as the outcome labels, respectively

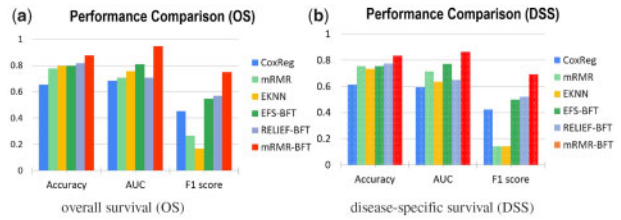

Fig. 4. Performance comparison of the proposed method and five other methods. The performance was evaluated by use of the matrices of Accuracy, AUC and F1 score and based on the outcome labels of OS (**a**) and DSS (**b**), respectively

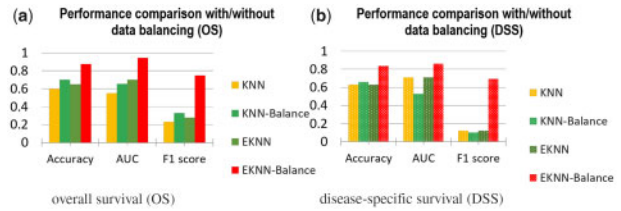

Fig. 5. Comparison of the proposed method with and without minor-class data balance. The classical K-NN method with and without data balance were compared as well

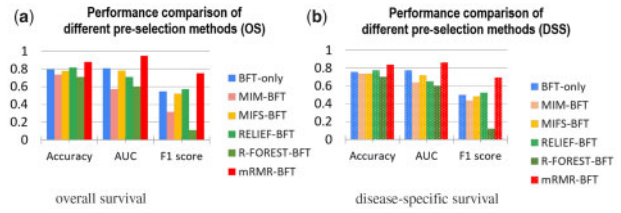

Fig. 6. Comparison of the mRMR-based feature pre-selection and other four feature pre-selection methods. The selected features from the five methods were employed to refine the features and the EK-NN classifier. BFT-only feature selection method was also compared. The performance was measured by use of the metrics of Accuracy, AUC and F1 score and based on the outcome labels of OS (**a**) and DSS (**b**), respectively

The effectiveness of the minor class data re-balance has been evaluated. Figure 5 showed the performance of the proposed method with and without data imbalance. In addition, the effectiveness of minor data re-balance was evaluated with the K-NN classifier. The results showed that minor class data re-balance improved the classifier performance on patient stratification. The effectiveness of the feature pre-selection process to improve the stratification performance of the learned EK-NN classifier was investigated. Figure 6 showed the prediction performance of the learned EK-NN prediction model by use the proposed mRMR-based method and four other pre-selection methods. These compared methods, which are independent to the classification performance, include RELIEF (Kira and Rendell, 1992a), mutual information maximization (MIM) (Fano and Hawkins, 1961) and mutual information feature selection (MIFS) (Battiti, 1994), and the random-forest method (Lin and Jeon, 2006). For fair comparison, a BFT-based feature refinement method was employed to refine the features pre-selected by these three methods, and the EK-NN-based classifier was employed as the pre-defined classifier for outcome prediction. The BFT method was directly applied to all profiled miRNA features to select a sparse subset to demonstrate the performance without any feature pre-selection. The parameters in all methods were optimized to achieve the best performance of each. The proposed mRMR-BFT method showed superior performance by selecting features considering both high feature-label relevance and low feature-feature redundancy.

The BFT-based feature refinement method was compared with other two widely used feature methods, the genetic algorithm (GA) (Davis, 1991) and the binary particle swarm optimization (BPSO) (Kennedy and Eberhart, 1997) methods. Both GA and BPSO methods require a pre-defined classifier for feature selection. Here, the mRMR-based selection method was uniformly employed as the feature pre-selection method for fair comparison. As shown in Figure 7, the proposed mRMR-BFT based method achieves superior prediction performance in terms of Accuracy, AUC and F1 score, compared to the other two methods.

The 5-year Kaplan–Meier survival curves were plotted for the prediction results of the proposed mRMR-BFT method and the other five compared methods (CoxReg, EFS-BFT, RELIEF-BFT, MIM-BFT and MIFS-BFT). The patients in the testing cohort were stratified into either the high-risk group or low-risk group based on the same threshold risk score determined by the training cohort of each of the compared methods. As shown in Figures 8 and 9, the proposed mRMR-BFT method can stratify high-risk and low-risk patient groups more accurately compared to the other five methods.

Table 2 shows the selected informative miRNA features by use of the proposed method and the other four methods: CoxReg, EFS-BFT, RELIEF-BFT and mRMR-only. In the CoxReg method (Gao *et al.*, 2013), a subset of 6 miRNA features were selected by use of multi-variate cox regression, which were employed to learn a cox regression model for stratifying patients into high-risk and low-risk groups. In the mRMR method, 13 miRNA features were selected and employed to train the EK-NN classifier directly without performing feature-refinement. A total of 5 and 8 features were selected by the EFT-BFT and RELIEF-BFT methods, respectively. In the proposed mRMR-BFT method, 13 miRNAs were pre-selected by the mRMR method, and then were refined to 6 features by use of the BFT-based refinement process through the optimization of the EK-NN classifier. It showed that less features were selected and the feature sparsity is high.

4 Discussion

In this study, a novel and systematic machine learning-based strategy was proposed to reliably stratify subsets of OPSCC patients with low and high risks of treatment failure. The proposed strategy included a two-stage feature selection procedure and an EK-NN classifier to address the challenges described above in Section 1. The model can serve as a clinical decision-making tool to (i) readily identify the subset of patients with low risk that would benefit from de-intensifying treatment, and (ii) accurately identify high risk patients for whom de-intensification would be detrimental and who may require further intensification. The designed method has several innovations. First, by use of appropriate methods to address the challenges in each step of outcome prediction, the overall performance is improved which is demonstrated by the prediction results when compared with other methods. The mass functions considered the possibilities of a sample belong to each single label and the

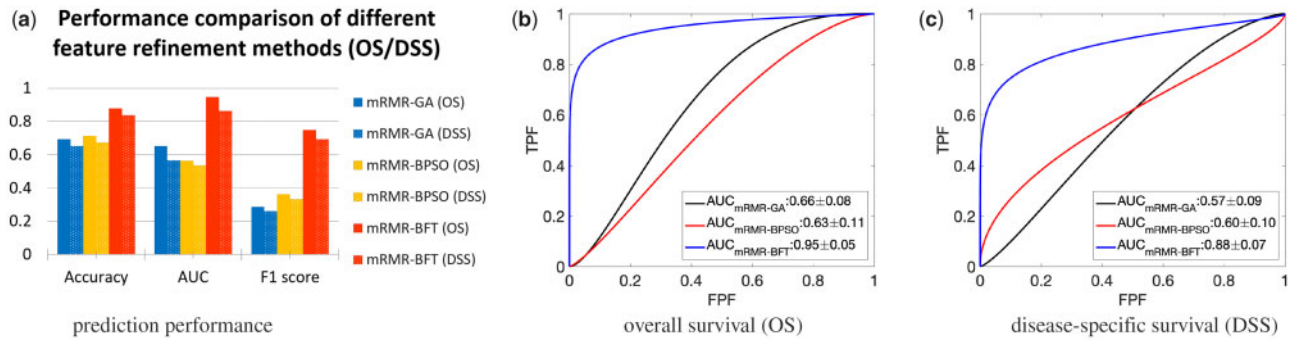

Fig. 7. Comparison of the BFT-based feature refinement method and two other feature refinement methods. The features selected by mRMR methods were employed as the input of three methods (**a**). The performance measured by use of metrics of Accuracy, AUC and F1 score (**a**). The ROC curves and corresponding AUC values by use of OS (b) and DSS (c) as the outcome labels

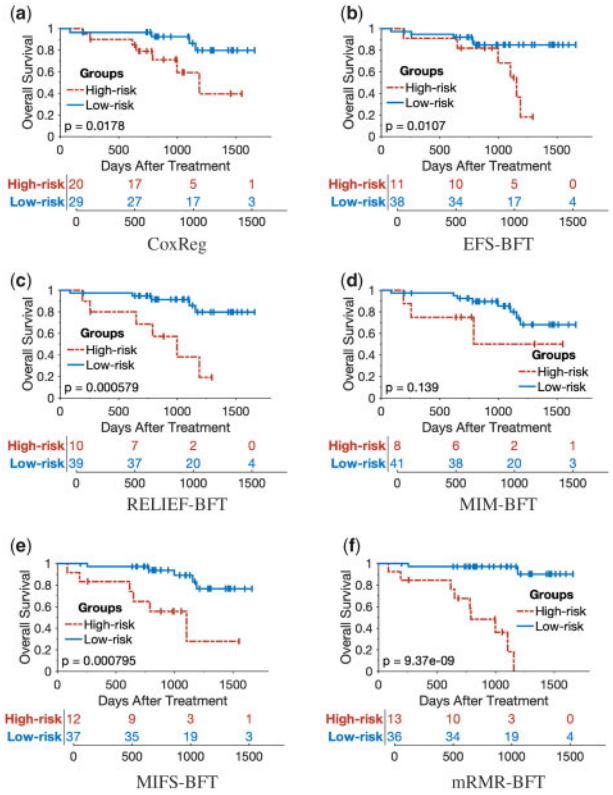

Fig. 8. Kaplan–Meier survival analysis to evaluate the performance of the proposed method and other five compared methods by use of OS as outcome labels

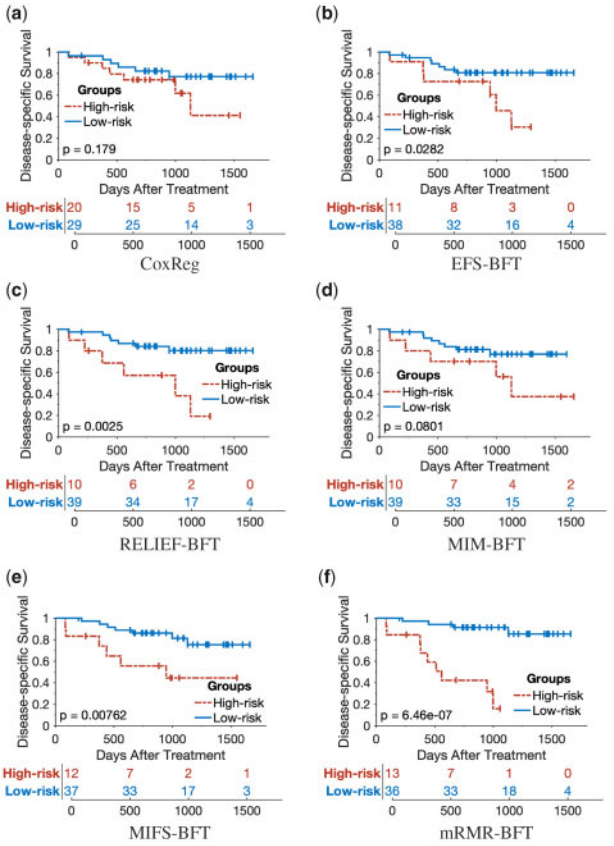

Fig. 9. Kaplan–Meier survival analysis to evaluate the performance of the proposed method and other five compared methods by use of DSS as outcome labels

combinations of multiple classes and addressed the uncertainty of features and labels of samples. In addition, the model was designed with a modularized structure, which facilitates the change of any one or several modules for classification performance evaluation/comparison and facilitate the incorporation of other state-of-the-art modules. The modularized design can provide a seamless integration of each component.

Some future research directions can be summarized as below. First, the proposed method was to train a prediction model for predicting oropharyngeal cancer treatment outcome and patient stratification. It forms a basis for biomarker and/or feature-based cancer treatment outcome prediction, and can be generalized to other type of cancer treatment outcome prediction scenarios. It was observed that the features selected (retained) from the total of 96 miRNA features show great variability in this study (Table 2). All these 96 miRNAs have been proved to be related to OPSCC outcomes. The selected feature subset relates to feature selection methods and classifiers. Further radiobiology investigation might be required to provide more information for future study. The proposed method should be further evaluated when larger clinical dataset is available. The performance improvement by using data-rebalancing process should be investigated with large datasets as well.

Second, radiomics, the high-throughput extraction and analysis of numerous features from medical images, is a highly promising approach for characterizing tumor phenotype, which provides an unprecedented opportunity to support and improve

Table 2. The miRNAs selected by varied methods

Methods	Selected miRNAs
CoxReg	miR-24, miR-31, miR-193b, miR-26b, miR-142-3p, miR-146a
mRMR-only	miR-31, miR-24, miR-215, miR-103, miR-26b, miR-25, miR-7, miR-148a, miR-30a-5p, miR-130a, miR-191, miR-16, miR-128b
EFS-BFT	miR-7, miR-9, miR-99a, miR-210, miR-220
RELIEF-BFT	miR-24, miR-31, miR-34c, miR-92, miR-135b, miR-210, miR-215, miR-328
mRMR-BFT	miR-7, miR-25, miR-31, miR-130a, miR-191, miR-215

personalized clinical decision-making (Wu *et al.*, 2019; Yip and Aerts, 2016) For several tumor sites, imaging biomarkers have shown promise in accurately separating favorable and unfavorable prognosis patients. Clinic features, such as gender, age and tumor stage, may also convey useful information for outcome prediction. However, current efforts to utilize high-dimensional multimodal biomarkers for early disease prognosis and treatment outcome prediction have also been compromised due to the above-mentioned challenges. Designing multimodal biomarker-based prognosis model can be potentially useful to learn a powerful and robust model for predicting treatment outcomes of oropharyngeal cancer and other cancer types. The thoughtful investigation of the correlation, independence and complementary nature of multimodal biomarkers (imaging, genomics, clinical and histopathologic biomarkers) remains unexplored, and requires further studies.

Third, deep learning methods have been applied in various fields and showed promising results. However, supervised deep learning methods have not been employed in this study, mainly because that the number of training samples is too small to well train a deep neural network, and data from single-modal has been employed, which will severely decrease the generalization ability of deep learning method. To study how to process multimodal data with deep learning method is a challenging task but an interesting research work that might provide more robust patient stratification and outcome prognosis. In addition, it will be interesting to assess the informativeness of the features extracted by use of deep learning methods and compare their performance with that of the features selected by traditional machine-learning methods.

5 Conclusion

A novel and systematic machine learning-based strategy was proposed to learn a prediction model with miRNA features for the stratification of oropharyngeal cancer patients with low and high risks of treatment failure. The prognosis model can be employed as a supporting tool to identify patients who are likely to fail standard therapy and potentially benefit from alternative or targeted treatments.

Funding

This work was supported by National Institutes of Health Award [R01DE026471, R01CA233873 and R21CA223799].

Conflict of Interest: none declared.

References

Ambros,V. (2004) The functions of animal MicroRNAs. *Nature*, **431**, 350–355.

Battiti,R. (1994) Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.*, **5**, 537–550.

Bland,J.M. and Altman,D.G. (1998) Survival probabilities (the Kaplan–Meier method). *BMJ*, **317**, 1572–1580.

Breiman,L. *et al.* (1984) *Classification and Regression Trees*. CRC Press, Boca Raton.

Chen,X. *et al.* (2018) Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics*, **34**, 4256–4265.

Cheng,G. *et al.* (2011) Conditional mutual information-based feature selection analyzing for synergy and redundancy. *ETRI J.*, **33**, 210–218.

Metz,E. (1999) *Rockit 0.9 b beta version*. IBM Compatible ROCKIT User's Guide. IBM, New York, NY, USA.

Cox,D.R. (1972) Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodological)*, **34**, 187–202.

Creed,J.H. *et al.* (2020) Matsurv: survival analysis and visualization in matlab. *J. Open Source Softw.*, **5**, 1830.

Damousis,I.G. *et al.* (2004) A solution to the unit-commitment problem using integer-coded genetic algorithm. *IEEE Trans. Power Syst.*, **19**, 1165–1172.

Davis,L. (1991) *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, NY, USA.

Dempster,A.P. (2008) Upper and lower probabilities induced by a multivalued mapping. In: Yager, R.R. and Liu L., (eds.) *Classic Works of the Dempster-Shafer Theory of Belief Functions. Studies in Fuzziness and Soft Computing*. Vol. 219. Springer, Berlin, Heidelberg. 10.1007/978-3-540-44792-4_3.

Denoeux,T. (1995) A k-nearest neighbor classification rule based on Dempster–Shafer theory. *IEEE Trans. Syst. Man Cybernet.*, **25**, 804–813.

,Denœux T. (2008) A k-nearest neighbor classification rule based on Dempster-Shafer theory. In: Yager R.R. and Liu L. (eds) Classic Works of the Dempster-Shafer Theory of Belief Functions. Studies in Fuzziness and Soft Computing. Vol 219. Springer, Berlin, Heidelberg. 10.1007/978-3-540-44792-4_29.

Denœux,T. and Kanjanatarakul,O. (2016) Evidential clustering: a review. In: *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM 2016),. Nov 2016, Da Nang, Vietnam*. pp. 24–35.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinf. Comput. Biol.*, **3**, 185–205.

Dudani,S.A. (1976) The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.*, **SMC-6**, 325–327.

Eid,H.F. *et al.* (2013) Linear correlation-based feature selection for network intrusion detection model. In: *International Conference on Security of Information and Communication Networks. 2013, 3-5 September; Cairo, Egypt*.

El Akadi,A. *et al.* (2011) A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge Inf. Syst.*, **26**, 487–500.

Ernster,J.A. *et al.* (2007) Rising incidence of oropharyngeal cancer and the role of oncogenic human papilloma virus. *The Laryngoscope*, **117**, 2115–2128.

Fano,R.M. and Hawkins,D. (1961) Transmission of information: a statistical theory of communications. *Am. J. Phys.*, **29**, 793–794.

Gao,G. *et al.* (2013) A MicroRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer*, **119**, 72–80.

Ge,R. *et al.* (2016) Mctwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics*, **17**, 142.

Gillison,M.L. *et al.* (2008) Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. *J. Natl. Cancer Inst.*, **100**, 407–420.

Hall,M.A. (2000) *Correlation-Based Feature Selection of Discrete and Numeric Class Machine Learning. Working paper Series, ISSN 1170-487X*.

Department of Computer Science, The University of Waikato, Private Bag 3105 Hamilton, New Zealand.
He,H. and Garcia,E.A. (2008) Learning from imbalanced data. *IEEE Trans. Knowledge Data Eng*., 1263–1284.
Heck,J.E. *et al.* (2010) Sexual behaviours and the risk of head and neck cancers: a pooled analysis in the international head and neck cancer epidemiology (INHANCE) consortium. *Int. J. Epidemiol*., **39**, 166–181.
Kennedy,J. (2006) Swarm intelligence. In: *Handbook of Nature-Inspired and Innovative Computing*. Springer, Boston, MA,pp. 187–219.
Kennedy,J. and Eberhart,R.C. (1997) A discrete binary version of the particle swarm algorithm. In: *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, **Vol. 5**. IEEE, pp. 4104–4108.
Kira,K. and Rendell,L.A. (1992a) The feature selection problem: traditional methods and a new algorithm. *AAAI*, **2**, 129–134.
Kira,K. and Rendell,L.A. (1992b) A practical approach to feature selection. In: *Machine Learning Proceedings 1992*. Morgan Kaufmann, pp. 249–256.
Kwak,N. and Choi,C.-H. (1999) Improved mutual information feature selector for neural networks in supervised learning. In: *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339), Vol. 2. IEEE, Glasgow, UK*, pp. 1313–1318.
Lian,C. *et al.* (2015) An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recogn*., **48**, 2318–2327.
Lian,C. *et al.* (2016a) Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction. *Med. Image Anal*., **32**, 257–268.
Lian,C. *et al.* (2016b) Robust cancer treatment outcome prediction dealing with small-sized and imbalanced data from FDG-PET images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Athens, Greece, pp. 61–69.
Lin,Y. and Jeon,Y. (2006) Random forests and adaptive nearest neighbors. *J. Am. Stat. Assoc*., **101**, 578–590.
Liu,Z. *et al.* (2017) Hybrid classification system for uncertain data. *IEEE Trans. Syst. Man Cybern. Syst*., **47**, 2783–2790.
Liu,Z. *et al.* (2018) Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Trans. Fuzzy Syst*., **26**, 1217–1230.
Loughrey,J. and Cunningham,P. (2004) Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, San Jose, California, USA, pp. 33–43.
Marur,S. and Burtness,B. (2014) Oropharyngeal squamous cell carcinoma treatment: current standards and future directions. *Curr. Opin. Oncol*., **26**, 252–258.
Masry,E. (1983) Probability density estimation from sampled data. *IEEE Trans. Inf. Theory*, **29**, 696–709.
Mi,H. *et al.* (2015) Robust feature selection to predict tumor treatment outcome. *Artif. Intell. Med*., **64**, 195–204.
Miller,D.L. *et al.* (2015) Identification of a human papillomavirus-associated oncogenic miRNA panel in human oropharyngeal squamous cell carcinoma validated by bioinformatics analysis of the cancer genome atlas. *Am. J. Pathol*., **185**, 679–692.
Nguyen,H.T. (2006) *An Introduction to Random Sets*. CRC Press.
Pearl,J. (1984) *Intelligent Search Strategies for Computer Problem Solving*. Addision Wesley Longman Publishing Co., Inc. 75 Arlington Street, Suite 300 Boston, MA, USA.
Peng,H. *et al.* (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell*., **28**, 1226–1238.
Quost,B. *et al.* (2011) Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules. *Int. J. Approx. Reason*., **52**, 353–374.
Ross,B.C. (2014) Mutual information between discrete and continuous data sets. *PLoS One*, **9**, e87357.
Satapathy,S. *et al.* (2017) MicroRNAs in HPV associated cancers: small players with big consequences. *Expert Rev. Mol. Diagn*., **17**, 711–722.
Shafer,G. (1976) *A Mathematical Theory of Evidence*, **Vol. 42**. Princeton University Press, Princeton, NJ, USA.
Steuer,R. *et al.* (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, S231–S240.
Street,W. (2019) *Cancer Facts & Figures 2019*. American Cancer Society, Atlanta, GA, USA.
Sun,Y. (2007) Iterative relief for feature weighting: algorithms, theories, and applications. *IEEE Trans. Pattern Anal. Mach. Intell*., **29**, 1035–1051.
Tan,M. *et al.* (2010) Learning sparse SVM for feature selection on very high dimensional datasets. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, June 21–24*, pp. 1047–1054.
Tang,J. *et al.* (2014) Feature selection for classification: a review. In: Data Classification: Algorithms and Applications, CRC Press, p. 37.
Walley,P. (2000) Towards a unified theory of imprecise probability. *Int. J. Approximate Reason*., **24**, 125–148.
Wang,T. *et al.* (2012) Correlation-based feature ordering for classification based on neural incremental attribute learning. *Int. J. Mach. Learn. Comput*., **2**, 807–811.
Wang,X. (2009) A PCR-based platform for MicroRNA expression profiling studies. *RNA*, **15**, 716–723.
Wen,T. *et al.* (2019) Maximal information coefficient-based two-stage feature selection method for railway condition monitoring. *IEEE Trans. Intell. Trans. Syst*., **20**, 2681–2690.
Winn,D. *et al.*; The INHANCE Consortium. (2015) The INHANCE consortium: toward a better understanding of the causes and mechanisms of head and neck cancer. *Oral Dis*., **21**, 685–693.
Wu,J. *et al.* (2019) Treatment outcome prediction for cancer patients based on radiomics and belief function theory. *IEEE Trans. Radiat. Plasma Med. Sci*., **3**, 216–224.
Yip,S.S. and Aerts,H.J. (2016) Applications and limitations of radiomics. *Phys. Med. Biol*., **61**, R150–R166.
Zouhal,L.M. and Denoeux,T. (1998) An evidence-theoretic k-nn rule with parameter optimization. *IEEE Trans. Syst. Man Cybernet. C (Appl. Rev.)*, **28**, 263–271.