A multi-objective radiomics model for the prediction of locoregional recurrence in head and neck squamous cell cancer

Kai Wang
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA

Zhiguo Zhou
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA
School of Computer Science and Mathematics, University of Central Missouri, Warrensburg, MO 64093, USA

Rongfang Wang
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA
School of Artificial Intelligence, Xidian University, Xi'an 710071, China

Liyuan Chen
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA

Qiongwen Zhang
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA
State Key Laboratory of Biotherapy and Cancer Center, Sichuan University and Collaborative Innovation Center, Chengdu 610041, China
Department of Head and Neck Cancer, West China Hospital, Chengdu 610041, China

David Sher and Jing Wang[a)]
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA

(Received 18 February 2020; revised 11 May 2020; accepted for publication 2 July 2020; published xx xxxx xxxx)

Purpose: Locoregional recurrence (LRR) is the predominant pattern of relapse after nonsurgical treatment of head and neck squamous cell cancer (HNSCC). Therefore, accurately identifying patients with HNSCC who are at high risk for LRR is important for optimizing personalized treatment plans. In this work, we developed a multi-classifier, multi-objective, and multi-modality (mCOM) radiomics-based outcome prediction model for HNSCC LRR.

Methods: In mCOM, we considered sensitivity and specificity simultaneously as the objectives to guide the model optimization. We used multiple classifiers, comprising support vector machine (SVM), discriminant analysis (DA), and logistic regression (LR), to build the model. We used features from multiple modalities as model inputs, comprising clinical parameters and radiomics feature extracted from X-ray computed tomography (CT) images and positron emission tomography (PET) images. We proposed a multi-task multi-objective immune algorithm (mTO) to train the mCOM model and used an evidential reasoning (ER)-based method to fuse the output probabilities from different classifiers and modalities in mCOM. We evaluated the effectiveness of the developed method using a retrospective public pretreatment HNSCC dataset downloaded from The Cancer Imaging Archive (TCIA). The input for our model included radiomics features extracted from pretreatment PET and CT using an open source radiomics software and clinical characteristics such as sex, age, stage, primary disease site, human papillomavirus (HPV) status, and treatment paradigm. In our experiment, 190 patients from two institutions were used for model training while the remaining 87 patients from the other two institutions were used for testing.

Results: When we built the predictive model using features from single modality, the multi-classifier (MC) models achieved better performance over the models built with the three base-classifiers individually. When we built the model using features from multiple modalities, the proposed method achieved area under the receiver operating characteristic curve (AUC) values of 0.76 for the radiomics-only model, and 0.77 for the model built with radiomics and clinical features, which is significantly higher than the AUCs of models built with single-modality features. The statistical analysis was performed using MATLAB software.

Conclusions: Comparisons with other methods demonstrated the efficiency of the mTO algorithm and the superior performance of the proposed mCOM model for predicting HNSCC LRR. [https://doi.org/10.1002/mp.14388]

Key words: head neck cancers, multi-objective model, outcome prediction, radiomics

1. INTRODUCTION

Head and neck squamous cell cancer (HNSCC) is one of the most common cancers worldwide.[1] Patients with HNSCC can be treated successfully in many cases, and radiation therapy is often required as part of managing their disease.[2,3] However, even after therapy with curative intent, 15% to 50% of patients with HNSCC will experience locoregional recurrence (LRR), mostly within 3 years after treatment.[4–6] Therefore, a strategy that can accurately identify patients at high risk for HNSCC LRR prior to treatment would help physicians to make better personalized treatment plans.

Radiomics-based methods have shown promising performance in HNSCC-related tasks, such as treatment outcome prediction, tumor segmentation, and pathologic classification.[7–12] These methods extract handcrafted quantitative features from radiological images, then use machine learning tools to analyze these features and build predictive models. Parmar *et al*. extracted radiomics features from pretreatment X-ray computed tomography (CT) images and built a Cox and logistic regression model for predicting local tumor control (LC) after chemoradiotherapy of HNSCC.[13] Their results demonstrated that some radiomics features were significantly associated with LC, and the best concordance index (CI) was 0.78. Vallieres et al. used a radiomics-based method to develop models that predict outcomes for LRR and distant metastases (DM) before HNSCC treatment.[11] They extracted radiomics features from 2-deoxy-2-[18F]fluoro-D-glucose positron emission tomography (FDG-PET) and CT scans and used a random forest method to construct a multivariable model. The best results were achieved when the model used features from multiple modalities, and they obtained area under the receiver operating characteristic curve (AUC) of 0.86 for DM and 0.69 for LRR.

One reason for the predictive capability of radiomics-based methods is that machine learning tools can learn complementary information from medical images using many different families of radiomics features.[9,14–16] This complementarity can stem from extracting different kinds of features using images from a single modality, extracting the same kind of feature from a single modality but under different scales, or jointly using different image and/or nonimage modalities. Furthermore, radiomics models that combine the output of different classifiers can achieve better accuracy and reliability than single-classifier–based models.[17] Multi-classifier (MC)-based approaches can further exploit the complementarity of radiomics features.

In addition to radiomics features, clinical variables such as age, tumor stage, disease site, and HPV status have the potential to improve the performance of LLR prediction model.[6,18–22] For patients with HNSCC in oropharynx, prospective clinical trial and retrospective analysis have shown that HPV status is strongly associated with therapeutic response and survival.[19–21] In the studies of HNSCC patient treatment response and survival prediction, age, tumor primary site, tumor T-stage, and tumor N-stage are found contributing significantly to overall survival.[6,18]

The design of the objective function is also of great importance when building radiomics models. Most state-of-the-art radiomics methods adopt AUC as the objective function to construct the predictive model. AUC summarizes the test performance over regions of the receiver operating characteristic curve (ROC) space and, thus, provides a measurement that accounts for both sensitivity and specificity. However, AUC is rarely used directly in clinical applications, and the final sensitivity and specificity are determined by the probability threshold, which needs to be manually selected according to clinical needs.[23–25] Take HNSCC LRR pretreatment prediction as an example: a model with high specificity is required to minimize false positives among high-risk patients receiving intensified therapy, which could lead to treatment-related toxicity.[26–28] By contrast, a highly sensitive model is preferred in other applications where additional treatment leads to negligible toxicity. To get a reliable prediction model, we developed a multi-objective model that considers both sensitivity and specificity simultaneously as the objective functions, and a preferred model can be selected from the Pareto-optimal solution set.[29]

In the present work, we have built a multi-classifier, multi-objective, and multi-modality (mCOM) model for HNSCC LRR prediction. Sensitivity and specificity were considered simultaneously as the objectives to guide the model construction, multiple classifiers were used to build the model, and both clinical features and radiomics features extracted from multiple modalities were used as model inputs. To solve the multi-objective optimization problem, we proposed a multi-task multi-objective immune algorithm (mTO) based on our previous work[29] to train the model.

2. MATERIALS AND METHODS

2.A. Data

2.A.1. Patients, images, and clinical parameters

We evaluated the proposed mCOM model on a publicly available HNSCC dataset downloaded from The Cancer Imaging Archive (TCIA).[11,30] This dataset consists of medical images and clinical data of 298 HNSCC patients from four different institutions: 91 from Hôpital général juif (HGJ) de Montréal, 101 from Centre hospitalier universitaire de Sherbrooke (CHUS), 41 from Hôpital Maisonneuve-Rosemont (HMR), and 65 from Centre hospitalier de l'Université de Montréal (CHUM). All these patients were treated with curative intent, 16% received radiation alone, and 84% received chemoradiation. The median time from pretreatment FDG-PET/CT imaging to first time of treatment was 18 days (range: 6–66 days), and the median follow-up duration after treatment for this study was 43 months (range: 6-112 months). Due to the retrospective nature of this dataset, imaging acquisition protocols were heterogeneous among different institutions. For example, the median PET acquisition time per bed position was 300 sec (range: 180–420 sec) and 150 sec (range: 120–151 sec) for patients in HGJ and CHUS,

respectively. Similarly, the median CT exposure was 210 mAs (range: 43–250 mAs) and 350 mAs (range: 5–350 mAs) for patients in the CHUM and CHUS, respectively. For each patient, gross tumor volumes (GTVs) were manually contoured on CT images by expert radiation oncologists and provided by the dataset. For 91 of the 298 patients (30.5%), GTVs are delineated on the CT of PET/CT and resampled to PET scan. For the other 207 patients (69.5%), the radiotherapy contours are drawn on the planning CT and then propagated/resampled to the PET/CT scan reference frame using intensity-based free-form deformable registration.[11] Forty-three (14.4%) of these patients experienced LRR. The median time to LRR after treatment for these patients was 18 months (range: 5–60 months), and six patients had LRR more than 3 years after the RT treatment. In this work, we did not exclude these patients with long interval between treatment and LRR in order to keep more minority cases (with LRR) and perform a direct comparison with results reported in the work by Vallieres et al. that included all patients with LRR. Extensive details regarding the patient cohort used in this work are publicly available on TCIA.[11,30]

After excluding patients whose weight and dose-related information in PET are incomplete and whose GTV segmentation is found not reasonable in the CT image when we reviewed the dataset, we used data from 277 patients in our study, 40 (14.4%) of whom experienced LRR. We also collected some clinical information, comprising age, sex, tumor T-stage, tumor N-stage, disease site, treatment paradigm, and HPV status to build the model. We selected these clinical variables based on strategies described by Vallieres et al.[11]

2.A.2. Radiomics feature extraction

We extracted a total of 257 features for each imaging modality using an open source radiomics software in the MATLAB environment.[31] These radiomics features consist of nine intensity features, eight geometry features, and 240 texture features. All features were extracted from the segmented 3D GTVs. For PET images, we calculated the standardized uptake value (SUV) before feature extraction.[32] For CT images, we kept their image values in raw Hounsfield unit (HU) format. As the slice thickness and in-plane resolution were different for PET and CT from different institutions, we used bilinear interpolation to make the images at the same resolution before feature extraction. Intensity features included minimum, maximum, mean, standard deviation, sum, median, skewness, kurtosis, and variance. Geometry features included volume, major diameter, minor diameter, eccentricity, elongation, orientation, bounding box volume, and perimeter. Texture features included energy, entropy, correlation, contrast, texture variance, sum-mean, inertia, cluster shade, cluster prominence, homogeneity, max-probability, and inverse variance. These texture features are based on 3D gray-level co-occurrence matrices (GLCMs) calculated using four different voxel distances (1 mm, 2 mm, 3 mm, and 4 mm) and five different gray levels (8, 16, 32, 64, and 128).

2.B. mCOM model

2.B.1. Overview

In mCOM, solutions are the basic predictive components that need to be trained and updated during model training, and some of them will be selected and used for prediction in testing. In this work, a solution is defined as the integration of different types of classifiers. These classifiers are trained with the same selected feature set, and their output probabilities are fused into a single value according to weighting factors to form one feasible solution. For a given modality, a solution is denoted by $\theta = \{f, \beta, w\}$, where f denotes the feature selection vector, β represents the parameters (including hyperparameters) of all the classifiers used in the solution, and w is the weighting factor of different classifiers for the classifier fusion. Different feasible solutions can be generated by changing the feature selection vector f, the hyperparameters of classifiers in β, or the weighting factor w.

The overall framework of the proposed mCOM model is shown in Fig. 1. The implementation of mCOM consists of a model training stage and a testing stage. The former uses training and validation data to generate a Pareto-optimal solution set, and the latter fuses the output probabilities of solutions in the Pareto-optimal solution set for testing samples and make the final prediction. In the training stage, after the solution set is initialized, the mTO algorithm optimizes the model through iterative feature selection, classifier parameter training, evidential reasoning (ER)-based fusion of output probabilities of classifiers, and Pareto-optimal solution set updating. After the model is well trained, the final output probabilities of testing samples are calculated in the testing stage by a two-step automatic weighted ER fusion method, which consists of solution fusion and modality fusion. The output probabilities of the Pareto-optimal solution set for each modality are fused first to calculate the output probability of that modality in solution fusion, then the output probabilities of all the modalities are fused to calculate the final output probability in modality fusion. Due to the population imbalance of LRR-positive and LRR-negative patients, synthetic minority oversampling technique (SMOTE)[33,34] is used to oversample the radiomics features and clinical parameters of LRR samples and produce class-balanced training data before the training stage.

2.B.2. Multiple objects in mCOM

Since the aim of HNSCC LRR outcome prediction is not only to obtain results with high accuracy, but also to get more reliable results with both high sensitivity and high specificity, a multi-objective model is desirable. To obtain a clinically desirable model, we simultaneously consider sensitivity and specificity as the objective functions for each modality, that is:

$$g = \max_{\theta} \left(g_{sen}, g_{spe} \right) \tag{1}$$

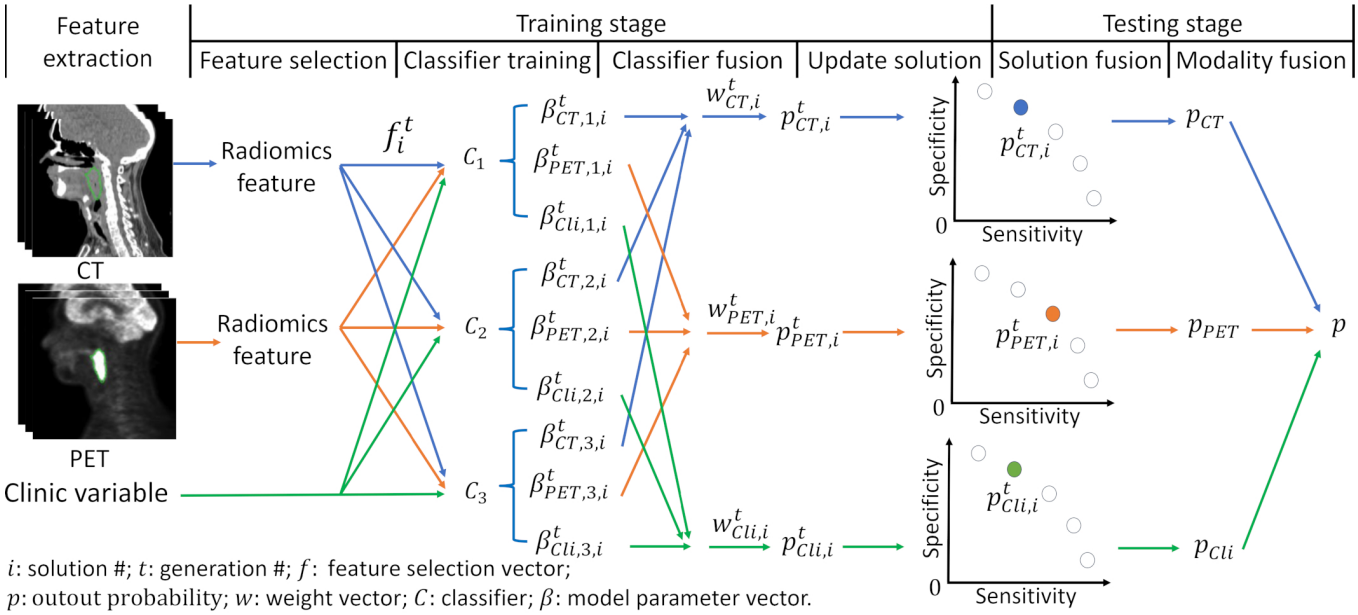

FIG. 1. The overall framework of the proposed model.

g_{sen} and g_{spe} are defined as:

$$g_{sen} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$g_{spe} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

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; 0.5 is used as the threshold for determining the predicted label according to the output probability. The goal of the proposed model is to maximize g_{sen} and g_{spe} simultaneously to obtain the Pareto-optimal solution set.

2.B.3. Multi-task multi-objective immune algorithm (mTO)

In our previous work,[29] we proposed an iterative multi-objective immune algorithm (IMIA) to solve the objective function, which adopts both sensitivity and specificity as optimization targets. Here, we propose mTO, which is a modified version of IMIA. There are two major differences between IMIA and mTO. First, because mCOM uses multiple classifiers to construct the predictive model, the weighting factors of these classifiers need to be optimized for classifier fusion through mTO. Second, in mCOM, the final output probability of each testing sample is obtained through solution fusion and modality fusion at testing stage, the weighting factors for these fusions are calculated in mTO, while IMIA relies on manually selecting a preferred solution from the Pareto-optimal solution set at the end of training for prediction.

mTO consists of six steps: initialization, cloning, mutation, deletion, solution update, and termination. In the initialization step, solution set S is randomly initialized as $S(0)$, and $S(0) = \{\theta_1, \ldots, \theta_{I_0}\}$, where I_0 is the number of solutions at the beginning of model training. Each individual solution $\theta_i, i = 1, 2, \ldots, I_0$, is defined as a group of parameters comprising feature selection vector f_i, classifier parameter vector β_i, and weighting factor vector w_i. f_i is a binary vector: a value of "1" indicates that the corresponding feature has been selected, while "0" indicates that it has not. β_i is the vector containing all the parameters, including hyperparameters, of different classifiers, and w_i is the weights used in classifier fusion to fuse the output probabilities of multiple classifiers into a single probability value.

After initialization, the first generation of solution sets for different modalities can be trained using features from training samples. A validation set is then used to evaluate all the solutions, and their performance—as measured by sensitivity, specificity, and AUC—is recorded and used as the basis for the solution cloning, mutation, and deletion operations in the next generation. In mTO, the cloning, mutation, deletion, and solution updating operations are the same as in IMIA; the detailed implementation can be found in our previous work.[29] When t reaches the maximal number of generations T, the weights for solution fusion and the weights for modality fusion are calculated based on the recorded performance of solutions on the validation data, and then the algorithm terminates.

2.B.4. Evidential reasoning

Among multi-modality radiomics methods, early integration and late integration are two typical approaches to fuse information provided by different modalities. Early integration methods directly concatenate all the features together to train a single model, while late integration methods construct separate classifiers using features from different modalities and combine the outputs of these classifiers using fusion

techniques. In our work, we used ER[35–37] as the fusion method for late integration of the output probabilities of individual classifiers or models. In the training stage, for each solution, ER fused the output probabilities of different classifiers; we defined this process as classifier fusion. In the testing stage, ER first fused the output probabilities of the Pareto-optimal solution set of each modality into one probability, then combined the probabilities from different modalities into one final output probability; we referred to these processes as solution fusion and modality fusion, respectively.

Take classifier fusion for a solution built with CT radiomics features as an example. Assume that, at generation t, $t = 1, 2, \ldots, T$ (where T is the number of the final generation), for a given feature selection vector $f^t_{CT,i}$, $i = 1, 2, \ldots, I$ (where I is the number of solutions at generation i), the probability vector $p = \{p^t_{CT,1,i}, p^t_{CT,2,i}, p^t_{CT,3,i}\}$ represents the output LRR probabilities from different classifiers C_1, C_2,C_3, and the parameters of these three classifiers are set as the elements in parameter vector $\beta^t_{CT,i} = \{\beta^t_{CT,1,i}, \beta^t_{CT,2,i}, \beta^t_{CT,3,i}\}$, separately. In this study, we chose three classifiers that are commonly used in radiomics: support vector machine (SVM), discriminant analysis (DA), and logistic regression (LR). Given the corresponding weight vector of this solution, $w^t_{CT,i} = \{w^t_{CT,1,i}, w^t_{CT,2,i}, w^t_{CT,3,i}\}$, which satisfies $\sum_{c=,1}^{3} w_{CT,c,i} = 1$, $0 \le w^t_{CT,c,i} \le 1$, the final output $p^t_{CT,i}$ is obtained through the following equations:

$$p^t_{CT,i} = \frac{\mu \times \left[\prod_{c=1}^{3}\left(w^t_{CT,c,i}p^t_{CT,c,i} + 1 - w^t_{CT,c,i}\right) - \prod_{c=1}^{3}\left(1 - w^t_{CT,c,i}\right)\right]}{1 - \mu \times \left[\prod_{c=1}^{C}\left(1 - w^t_{CT,c,i}\right)\right]} \quad (4)$$

where μ is calculated as:

$$\mu = \left[\prod_{c=1}^{3}\left(w^t_{CT,c,i}p^t_{CT,c,i} + 1 - w^t_{CT,c,i}\right) + \prod_{c=1}^{3}\left(1 - w^t_{CT,c,i}p^t_{CT,c,i}\right) - \prod_{c=1}^{3}\left(1 - w^t_{CT,c,i}\right)\right]^{-1} \quad (5)$$

In the training stage, the weighting factor w for classifier fusion is generated and iteratively updated by the mTO algorithm. In the testing stage, weighting factor w' for solution fusion and modality fusion is calculated according to a weighting function that can tune the balance between model sensitivity and specificity. In this study, we wish to get a balanced result, and we set the weight for solutions with extremely imbalanced sensitivity or specificity to zero. The weighting function is defined as:

$$w' = \begin{cases} \frac{g_{sen}}{g_{spe}} + AUC & \text{when} 0.5 \le \frac{g_{sen}}{g_{spe}} \le 1 \\ \frac{g_{spe}}{g_{sen}} + AUC & \text{when} 0.5 \le \frac{g_{spe}}{g_{sen}} \le 1 \\ 0 & \text{Other situations} \end{cases}. \quad (6)$$

where g_{sen}, g_{spe}, and AUC denote the sensitivity, specificity, and AUC values, respectively, of each solution on the validation set, and w' is normalized into a unit vector thereafter. With the weighting vector automatically generated according to the performance on validation data, the ER fusion method can first fuse the output probabilities of solutions in the Pareto-optimal set for each modality in solution fusion, and then fuse the output probabilities of each modality into the final probability in modality fusion.

2.B.5. Testing procedure of the mCOM model

After the mTO terminates, the Pareto-optimal solution set of each modality is well trained on the training and validation data. The solution for each modality S_m is the fusion of solutions in the corresponding Pareto-optimal set, which is denoted as:

$$S_m = \mathrm{ER}(\theta_1, \ldots, \theta_{I_{max}}, w_{sol}) \quad (7)$$

where $\theta_1, \ldots, \theta_{I_{max}}$ are the solutions selected from the Pareto-optimal set after T generations, I_{max} is the number of these solutions, and w_{sol} is the automatically generated weighting factor vector for solution fusion; this vector is calculated using Eq. (6) based on the validation performance of each solution. The final solution for the whole model is defined as:

$$S_{fin} = \mathrm{ER}(S_1, \ldots, S_M, w_{mod}), \quad (8)$$

where $S_1, \ldots, S_M$ are the fused solutions for different modalities, M is the number of modalities (which is 3 in this study), and w_{mod} is the automatically generated weighting factor vector for modality fusion; this vector is calculated using Eq. (6) based on the validation performance of each modality. For a testing sample, first, the features for each solution in S_m (m = 1, 2, 3) are selected; second, each solution in S_m outputs a probability $p^T_{m,i}$ (i = 1, 2,…, I_{max}) through its internal classifier fusion; third, the output probability of the solution set for each modality p_m is obtained through solution fusion of $p^T_{m,i}$ and p_m to yield the final output probability for each single-modality model; finally, modality fusion fuses the p_m of different modalities to calculate the final output probability of the multi-modality model, and the label can be determined then.

3. RESULTS

3.A. Experimental setup

In the training stage of mCOM, the population number I was set to 100, and the maximal generation number T was set to 50. In the cloning step, the expected value for the clonal solution set n_{clo} was set to 200. In the mutation step, the mutation probability MP was set to 0.9. To facilitate the model training process, we used the minimal-redundancy-maximal-relevance criterion (mRMR) method,[38] a standard approach that is widely used for feature selection,[39–41] to preselect features from the radiomics features extracted from PET and CT, and we set the number of preselected features to 50.

We designed three groups of experiments. (a) To show the benefits of adopting multiple classifiers (MC), we compared

the performances of single-classifier models built from each base classifier individually (with all other settings kept the same as mCOM) with the mCOM model's performance. (b) To show the benefits of adopting features from multiple modalities, we compared the performances of models built using features from different modalities, both individually and in various combinations. (c) To show the benefits of automatic weighted fusion, we evaluated the robustness of the mCOM model's fusion strategy in solution fusion and modality fusion and compared it with a manually selected single solution.[29]

We used sensitivity, specificity, accuracy, ROC curve, and AUC to evaluate the performance of the different models. We also investigated the statistical significance of the differences between AUCs of different models using paired z-test with a significance level of 0.05.[42] To make a direct and fair comparison with the results reported by Vallieres et al., for TCIA HNSCC dataset, fixed training and testing samples were adopted as they described.[11] Data from patients treated at HGJ and CHUS were used as the training and validation sets. During the training process, four-fifths of the data from HGJ and CHUS were randomly selected as the training set, and the rest were used as the validation set. Data of patients treated at HMR and CHUM were used as the testing set. To reduce variability caused by subset partition in the training stage, all experiments were performed five times. The mean and standard deviation was then calculated for each evaluation criterion. In the remainder of the paper, all prediction results reported were from the testing set.

3.B. Performance evaluation of multiple classifiers vs single classifiers

The prediction results of different classifiers using clinical data, CT, and PET pretreatment radiomics features are shown in Table I; the *p*-values were calculated between the MC model and the single-modality models for each modality. For clinical data, the single-classifier models and the MC model yielded very similar results. For models trained with CT radiomics features, the fusion of different classifiers yielded better results than single classifiers on all evaluation metrics, although the AUC was not significantly higher than with single-classifier models. For models trained with PET radiomics features, the MC model achieved a significantly higher AUC value than models based on SVM and LR.

3.C. Performance evaluation of multiple modalities vs single modalities

The evaluation results on single and multiple modalities for pretreatment LRR prediction are shown in Table II and Fig. 2; the *p*-values are calculated from the comparison between the multi-modality model built with CT, PET, and clinical features, the multi-modality model built with just CT and PET features, and the single-modality models built with CT, PET, and clinical features individually. The AUCs of the multi-modality models (0.76 for the fusion of CT and PET radiomics features, and 0.77 for the fusion of CT and PET radiomics features and clinical parameters) were more than 9.54% higher than those of the single-modality models. According to the *p*-values, the fusion of all the three modalities yielded a significantly higher AUC than other models.

Our mCOM model achieved higher AUC values, for both the radiomics-only model and the model built with radiomics and clinical features, than the multi-modality models in Vallieres et al.[11] For the radiomics-only model, our mCOM model achieved an AUC of 0.76, while Vallieres' model achieved an AUC of 0.64. Similarly, for the model built with all three modalities, our mCOM model achieved an AUC of 0.77, while Vallieres' model obtained an AUC of 0.69.

3.D. Performance evaluation of automatic weighted fusion vs manual selection

The comparisons between the final solution manually selected according to our previous method[29] and the automatic weighted solution fusion in mCOM are shown in Table III; the *p*-values were calculated between these two methods.

TABLE 1. Performance of models built with different classifiers and features from different modalities.

Modality	Classifier	Sensitivity	Specificity	Accuracy	AUC	*P*-value
Clinic	SVM	0.60 ± 0.20	0.67 ± 0.22	0.66 ± 0.16	0.66 ± 0.06	0.51
	LR	0.65 ± 0.07	0.63 ± 0.06	0.63 ± 0.04	0.68 ± 0.02	0.98
	DA	0.62 ± 0.08	0.63 ± 0.05	0.63 ± 0.04	0.67 ± 0.01	0.52
	MC	0.63 ± 0.06	0.65 ± 0.06	0.64 ± 0.06	0.68 ± 0.02	–
CT	SVM	0.52 ± 0.06	0.82 ± 0.08	0.78 ± 0.06	0.66 ± 0.01	0.89
	LR	0.52 ± 0.03	0.84 ± 0.01	0.79 ± 0.01	0.67 ± 0.02	0.59
	DA	0.54 ± 0.00	0.81 ± 0.02	0.77 ± 0.02	0.67 ± 0.01	0.87
	MC	0.54 ± 0.00	0.84 ± 0.02	0.80 ± 0.01	0.69 ± 0.02	–
PET	SVM	0.62 ± 0.09	0.49 ± 0.04	0.51 ± 0.02	0.52 ± 0.01	<0.01
	LR	0.59 ± 0.09	0.50 ± 0.04	0.51 ± 0.03	0.53 ± 0.01	0.02
	DA	0.58 ± 0.14	0.61 ± 0.03	0.61 ± 0.02	0.59 ± 0.03	0.26
	MC	0.62 ± 0.12	0.61 ± 0.03	0.61 ± 0.01	0.62 ± 0.03	–

DA, discriminant analysis, LR, logistic regression; MC, multi-classifier model; SVM, The classifiers comprise support vector machine.

TABLE II. Performance of multi-classifier models (MCs) built with different combinations of modalities.

Modality	Classifier	Sensitivity	Specificity	Accuracy	AUC	*P*-value
Clinic	MC	0.63 ± 0.06	0.65 ± 0.06	0.64 ± 0.06	0.68 ± 0.02	0.02
CT	MC	0.54 ± 0.00	0.84 ± 0.02	0.80 ± 0.01	0.69 ± 0.02	<0.01
PET	MC	0.62 ± 0.12	0.61 ± 0.03	0.61 ± 0.01	0.62 ± 0.03	<0.01
CT + PET	MC	0.54 ± 0.00	0.85 ± 0.01	0.80 ± 0.01	0.76 ± 0.00	0.02
CT + PET+Clinic	MC	0.54 ± 0.00	0.84 ± 0.02	0.80 ± 0.02	0.77 ± 0.00	–

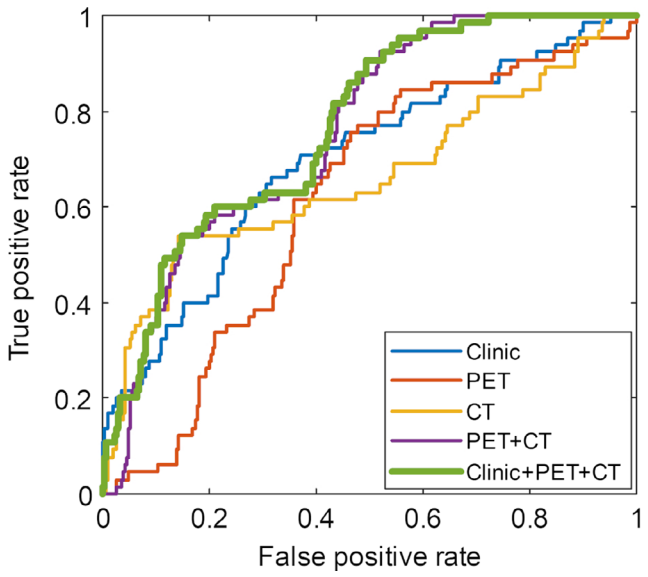

FIG. 2. Receiver operating characteristic curves (ROCs) of models built with features from different modalities.

For manual selection, the best solution for each modality was selected according to the performance on the validation set. For automatic weighted fusion, the fused solution was calculated as described in Section 2.B.5. The automatic weighted fusion method yielded better average performance and lower variance on all evaluation metrics, and the AUC value was significantly higher than the manual selection method.

4. DISCUSSION

The motivation of this study was to develop a method that employs radiomics-based approaches to predict HNSCC LRR prior to treatment. In our previous study, we identified several advantages of our multi-objective, multi-modality, and multi-classifier radiomics model for some outcome prediction applications, such as predicting DM in HNSCC, non-small cell lung cancer, and cervical cancer.[12,29,43,44] However, the previous approaches focused on only one or two of the aspects in the current mCOM, such as a multi-objective model trained with features from a single modality,[44] or a multi-modality model built with a single classifier.[29] In addition, these approaches also required that the best solution be selected manually from the Pareto-optimal solution set for the final prediction. This study still follows the same general framework, as it simultaneously uses sensitivity and specificity as the objective functions, and it uses the ER method to fuse output probabilities of different classifiers or solutions during training. However, instead of manually selecting a single solution for testing, we specifically modified the IMIA algorithm to train the model using a weighting function to fuse all the solutions in the Pareto-optimal set. Furthermore, we integrated the concepts of multi-objective, multi-modality, and multi-classifier models together into a unified mCOM framework. We found these approaches to be effective in our experiments for HNSCC LRR prediction.

For HNSCC LRR prediction, models built with multiple classifiers yielded better AUC values than those built with single classifiers, which indicates that the MC strategy could achieve more robust results, although the differences were not significant for some modalities. For the multi-modality strategy, fusing models for different modalities improved the performance significantly. Compared with the models proposed by Vallieres et al,[11] our mCOM model yielded an AUC that was 12.76% higher for the radiomics-only model and 11.14% higher for the model built with all three modalities.

In the training stage, the goal of the mCOM method is to find the Pareto-optimal solution set for each modality. All the solutions in these solution sets are fused together through a two-step automatic weighted fusion in the testing stage. Since the number of samples is often limited in treatment outcome prediction applications and the populations of positive and negative cases are imbalanced, manually selecting the best solution from the Pareto-optimal set according to the performance on one validation set may not yield an optimal result.

TABLE III. Results of manual selection and automatic weighted solution fusion in the proposed multi-classifier, multi-objective, and multi-modality model (mCOM).

Method	Sensitivity	Specificity	Accuracy	AUC	*P*-value
Manual	0.52 ± 0.03	0.82 ± 0.08	0.77 ± 0.07	0.74 ± 0.02	<0.01
Automatic	0.54 ± 0.00	0.84 ± 0.02	0.80 ± 0.02	0.77 ± 0.00	

According to Table III, the proposed automatic weighted fusion method is more stable than selecting a preferred solution manually.

In this study, we used a balanced weighting function for automatic weighted fusion in expectation of similar sensitivity and specificity on the testing data. However, the sensitivity and specificity of mCOM—0.54 and 0.84, respectively—on the test data are not balanced. One possible explanation, also raised by Vallieres et al,[11] is that this imbalance might be caused by the patient population differences between the training and testing samples. The data utilized as training and validation are from two institutions, and the testing data are from two other institutions. Nevertheless, the preference for higher sensitivity or higher specificity could be achieved flexibly by increasing the weight for different objectives when fusing the solutions, or by manually selecting the output probability threshold in determining positive or negative predicted outcome. On the other hand, for HNSCC LRR pretreatment prediction, treatment intensification might be applied to patients who are predicted to have higher risk of LRR, and this intensified therapy could lead to treatment-related toxicity. In this context, our high specificity model is suitable to avoid false positives among high-risk patients receiving intensified therapy.

There are several potential improvements of the current study worth mentioning. First, as the number of patients with LRR in the test cohort was already small, we did not construct and test models based on disease subsites. It is worthy to analyze the performance of treatment outcome prediction for HPV-associated oropharyngeal cancers vs other HPV-negative head and neck cancers in a larger patient cohort. Second, three simple classifiers were used together in mCOM, and the improvement on AUC is not significant comparing with single-classifier models using clinical feature and CT. With the development of machine learning, there are many different types of classifiers available. More classifiers could be integrated into our model in future works to achieve better model robustness and reliability. Third, the features utilized in this study are all handcrafted. Since deep learning methods like convolution neural network (CNN) have shown great success in image processing applications and can extract high-level features through the model training process, the features from handcrafted methods and learning-based methods, for example features learned from CNN, could be utilized together to build the predictive model. Fourth, as the dataset used in this work was collected retrospectively, various protocols were used for both CT and PET image acquisition. As radiomics features are affected by image acquisition protocols, the performance of our model may be affected by inhomogeneous image acquisition protocols. Protocol standardization for image acquisition is desired in prospective studies to minimize the influence of protocol variations. Finally, genetic testing is rapidly becoming an important tool in the management of patients with HNSCC, and genetic factors have been found to be crucial in tumor progression and response to treatment.[45–47] In our future work, we will construct an outcome prediction model that uses genetic information as an additional modality to further increase the information complementarity.

5. CONCLUSIONS

We proposed a mCOM radiomics model to predict HNSCC LRR. SVM, LR, and DA were used to build the model, sensitivity and specificity were simultaneously considered as objectives to guide the model optimization, and CT and PET radiomics features and clinical parameters were used as model inputs. Using mCOM, we achieved AUC values above 0.76 on TCIA pretreatment HNSCC dataset. Fusing multiple classifiers and multiple modalities can improve the robustness of the predictive model. Additionally, the multi-objective model training method together with the automatic weighted fusion strategy could help to build a predictive model that can be flexibly adapted to different clinical preferences. The proposed mCOM model can be applied not only to HNSCC LRR prediction, but also to other outcome prediction problems in medicine.

ACKNOWLEDGMENTS

This work is supported by the Cancer Prevention and Research Institute of Texas (RP160661) and National Institutes of Health (R01 EB027898).

CONFLICT OF INTEREST

The authors have no relevant conflict of interest to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: jing.wang@utsouthwestern.edu.

REFERENCES

1. Sanderson RJ, Ironside JA. Squamous cell carcinomas of the head and neck. *BMJ*. 2002;325:822–827.
2. Bourhis J, Overgaard J, Audry H, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. *Lancet*. 2006;368:843–854.
3. Ratko TA, Douglas GW, de Souza JA, Belinson SE, Aronson N. *Radiotherapy Treatments for Head and Neck Cancer Update*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2014.
4. Beswick DM, Gooding WE, Johnson JT, Branstetter BF. Temporal patterns of head and neck squamous cell carcinoma recurrence with positron-emission tomography/computed tomography monitoring. *Laryngoscope*. 2012;122:1512–1517.
5. Oksuz DC, Prestwich RJ, Carey B, et al. Recurrence patterns of locally advanced head and neck squamous cell carcinoma after 3D conformal (chemo)-radiotherapy. *Radiat Oncol*. 2011;6:54.
6. Chang JH, Wu CC, Yuan KS, Wu ATH, Wu SY. Locoregionally recurrent head and neck squamous cell carcinoma: incidence, survival, prognostic factors, and treatment outcomes. *Oncotarget*. 2017;8:55600–55612.
7. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging*. 2010;10:137–143.
8. Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. *Transl Cancer Res*. 2016;5:371–382.

9. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–577.
10. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14:749–762.
11. Vallieres M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep-Uk*. 2017;7.
12. Chen LY, Zhou ZG, Sher D, et al. Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Phys Med Biol*. 2019;64:075011.
13. Parmar C, Leijenaar RTH, Grossmann P, et al. Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer. *Sci Rep-Uk*. 2015;5.
14. Cao HL, Bernard S, Sabourin R, Heutte L. Random forest dissimilarity based multi-view learning for Radiomics application. *Pattern Recogn*. 2019;88:185–197.
15. Fave X, Zhang LF, Yang JZ, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep-Uk*. 2017;7.
16. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ Breast Cancer*. 2017;3:43.
17. Chen X, Zhou Z, Hannan R, et al. Reliable gene mutation prediction in clear cell renal cell carcinoma through multi-classifier multi-objective radiogenomics model. *Phys Med Biol*. 2018;63:215008.
18. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*. 2001;23:718–724.
19. Beesley LJ, Hawkins PG, Amlani LM, et al. Individualized survival prediction for patients with oropharyngeal cancer in the human papillomavirus era. *Cancer*. 2019;125:68–78.
20. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010;363:24–35.
21. Fakhry C, Westra WH, Li S, et al. Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. *J National Cancer Institute*. 2008;100:261–269.
22. Goon PK, Stanley MA, Ebmeyer J, et al. HPV & head and neck cancer: a descriptive update. *Head Neck Oncol*. 2009;1:36.
23. Freeman EA, Moisen GG. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*. 2008;217:48–58.
24. Chen J, Tsai C-A, Moon H, Ahn H, Young J, Chen C-H. Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*. 2006;17:337–352.
25. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeogra*. 2008;17:145–151.
26. Brockstein BE, Vokes EE. Head and neck cancer in 2010: maximizing survival and minimizing toxicity. *Nat Rev Clin Oncol*. 2011;8:72–74.
27. Braaksma M, van Agthoven M, Nijdam W, Uyl-de Groot C, Levendag P. Costs of treatment intensification for head and neck cancer: concomitant chemoradiation randomised for radioprotection with amifostine. *Eur J Cancer*. 2005;41:2102–2111.
28. Trotti A. Toxicity in head and neck cancer: a review of trends and issues. *Int J Radiat Oncol Biol Phys*. 2000;47:1–12.
29. Zhou ZG, Folkert M, Iyengar P, et al. Multi-objective radiomics model for predicting distant failure in lung SBRT. *Phys Med Biol*. 2017;62:4460–4478.
30. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045–1057.
31. Yang X, Tridandapani S, Beitler JJ, et al. Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: an in vivo study of late toxicity. *Med Phys*. 2012;39:5732–5739.
32. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR Am J Roentgenol*. 2010;195:310–320.
33. Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*. 2013;14:106.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artificial Intelligence Res*. 2002;16:321–357.
35. Yang JB, Xu DL. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *Ieee T Syst Man Cy A*. 2002;32:289–304.
36. Wang YM, Yang JB, Xu DL. Environmental impact assessment using the evidential reasoning approach. *Eur J Oper Res*. 2006;174:1885–1913.
37. Yang JB, Xu DL. Evidential reasoning rule for evidence combination. *Artif Intell*. 2013;205:1–29.
38. Peng HC, Long FH, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *Ieee T Pattern Anal*. 2005;27:1226–1238.
39. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol*. 2015;114:345–350.
40. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272.
41. Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino acids*. 2012;42:1387–1395.
42. Zhang DD, Zhou XH, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Stat Med*. 2002;21:701–715.
43. Zhou Z, Wang K, Liu H, Sher D, Wang J. Multifaceted Radiomics: towards more reliable radiomics for predicting distant metastasis in head & neck cancer. *Med Phys*. 2019;46:E404–E404.
44. Hao HX, Zhou ZG, Li SL, et al. Shell feature: a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer. *Phys Med Biol*. 2018;63:095007.
45. Lacko M, Braakhuis BJ, Sturgis EM, et al. Genetic susceptibility to head and neck squamous cell carcinoma. *Int J Radiat Oncol Biol Phys*. 2014;89:38–48.
46. Birkeland AC, Uhlmann WR, Brenner JC, Shuman AG. Getting personal: Head and neck cancer management in the era of genomic medicine. *Head & Neck*. 2016;38:E2250–E2258.
47. Vossen DM, Verhagen CVM, van der Heijden M, et al. Genetic factors associated with a poor outcome in head and neck cancer patients receiving definitive chemoradiotherapy. *Cancers (Basel)*. 2019;11:445.