Oropharyngeal Cancer Patient Stratification Using Random Forest Based-learning Over High-dimensional Radiomic Features

Harsh Patel
University of Iowa
David Vock
University of Minnesota
Elisabeta Marai
University of Illinois at Chicago
Clifton Fuller
The University of Texas MD Anderson Cancer Center
Abdallah Mohamed
The University of Texas MD Anderson Cancer Center
Guadalupe Canahuate (guadalupe-canahuate@uiowa.edu)
Department of Electrical and Computer Engineering, University of Iowa, Iowa City, USA

Research Article

Keywords: Cluster Analyses, Dimensionality Reduction, Radiomics, Random Forest Clustering, Head and Neck Cancer

Posted Date: January 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-135236/v1

Version of Record: A version of this preprint was published at Scientific Reports on July 7th, 2021. See the published version at https://doi.org/10.1038/s41598-021-92072-8.

Oropharyngeal cancer patient stratification using random forest based-learning over high-dimensional radiomic features

Harsh Patel[1], David M. Vock[2], G. Elisabeta Marai[3], Clifton D. Fuller[4], Abdallah S. R. Mohamed[4], Guadalupe Canahuate[1*]

[1]*Department of Electrical and Computer Engineering, University of Iowa, Iowa City, USA*
[2]*Division of Biostatistics, University of Minnesota, Minneapolis, USA*
[3]*Department of Department of Computer Science, University of Illinois at Chicago, Chicago USA*
[4]*Department of Radiation Oncology, MD Anderson Cancer Center, Houston, USA*

Keywords: Cluster Analyses, Dimensionality Reduction, Radiomics, Random Forest Clustering, Head and Neck Cancer
Word Count: 3879

Abstract

OBJECTIVE: To improve risk prediction for oropharyngeal cancer (OPC) patients using cluster analysis on the radiomic features extracted from pre-treatment Computed Tomography (CT) scans.

MATERIALS AND METHODS: OPC Patients were classified into 2 or 3 risk groups by applying hierarchical clustering over the co-occurrence matrix obtained from a random survival forest (RSF) trained over 301 radiomic features. The cluster label was included together with other clinical data to train an ensemble model using five predictive models (Cox, random forest, RSF, logistic regression, and logistic-elastic net). Ensemble performance was evaluated over an independent test set for both recurrence free survival (RFS) and overall survival (OS).

RESULTS: The Kaplan-Meier curves for OS stratified by cluster label show significant differences for both training (p-val<0.0001) and testing (p-val=0.005). Inclusion of the cluster label outperforms clinical data only improving AUC from .60 to .76 and from .63 to .75 for OS and RFS, respectively.

CONCLUSION: The extraction of a single feature, namely a cluster label, to represent the high-dimensional radiomic feature space reduces the dimensionality and sparsity of the data. Moreover, inclusion of the cluster label improves model performance compared to clinical data only and compares to the raw radiomic features performance.

Background and Significance

Radiomics entails extraction of quantitative imaging features from computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) images. A large number of radiomic features can be extracted from these images to characterize tumor intensity, shape, and texture. Dimensionality reduction can significantly reduce the number of features which represent the high-dimensional radiomic space. Dimensionality reduction seeks to identify tumor signature profiles that can be used for prognostic or predictive evaluation of patient outcomes [1, 50], and have been putatively associated with clinical and survival outcomes [2, 3, 4, 5].

Dimensionality reduction can be categorized into either feature selection or feature extraction. With feature selection, a subset of the original feature set is used for prediction and focuses on a number of studies dealing with radiomic data [6, 7, 8]. The feature set is transformed into a new smaller set of features that captures much of the high-dimensional space variance with feature extraction. Furthermore, feature extraction finds a new representation of the features which more concisely represent the entire feature space. Feature clustering is an approach to feature extraction used to reduce radiomics dimensionality [6, 9, 10]. Clustering used as a feature extraction method can represent an entire set of radiomic features and benefit from massively reducing the radiomic feature space into a single covariate. The cluster label also allows easy visualization and differentiation of the patients, which is difficult with feature selection alone.

In supervised dimensionality reduction, the outcome of interest is considered when producing a radiomic signature for either feature extraction or selection. Some studies have examined the use of unsupervised methods for event prediction with radiomic data [11, 12], but

the inclusion of an outcome in the dimension reduction process has the potential to increase predictive power.

Survival endpoints, such as overall survival (OS), local recurrence control (LC), distant metastasis (DM), regional recurrence control (RC), or combined outcomes such as recurrence free survival (RFS) are considered right-censored when the time-to-event is unknown at the end of an individual's follow-up. That is, at any given point during follow-up, some patients are yet without an event but still potentially at risk for an event with further follow-up. Samples for which the outcome has not been observed at the last follow up are said to be right-censored. Several standard machine learning applications have been extended to allow the use of right-censored data [16]. Some methods (e.g., random survival forests) have been developed to perform feature selection using the right-censored outcomes directly; that is, these methods directly account for the unequal follow-up time among individuals [7].

Objective

This paper focuses on developing a novel methodology for feature extraction from a high-dimensional set of radiomic features using random survival forest to cluster the patients and use the cluster label in posterior analyses to represent the entire radiomic feature space. Random forest (RF) is an increasingly popular approach for dealing with high dimensional data. A random forest is an ensemble-based decision tree method used for classification and feature selection. Random forests have been adapted to extend beyond a categorical outcome; random survival forests (RSF) [15] use the right-censored outcome directly. Specifically, we propose using the proportion of times a pair of patients fall into the same terminal nodes in the trees of the random forest as a similarity metric to cluster the patients. This method is known as random forest clustering [18], but previous studies [18, 19, 20] have used random forest clustering for

unsupervised learning to cluster unlabeled data. Our work differs from this previous work in that we are applying this to already labeled survival data to extract a single covariate, which can then be used to build predictive models. We use selected features and a trained regression model to assign previously unseen test samples into a cluster. Subsequently, the cluster label is used as a covariate for risk prediction from an ensemble model of established risk prediction approaches (Cox Proportional Hazard, Random Forest, Random Survival Forest, Logistic Regression, and Logistic-Elastic Net), which have been adapted to right-censored outcomes using inverse probability of censoring weights [17].

Materials and Methods

Data Source

Our institutional database was retrospectively reviewed for oropharyngeal cancer patients treated at MD Anderson Cancer Center during the period of (2005-2013) following Institutional Review Board (IRB) approval. Eligible patients diagnosed with oropharyngeal cancers were pathologically confirmed either by a biopsy or a surgical excision and received their treatment (i.e., chemo-radiotherapy) with curative intent.

For imaging data, contrast-enhanced computed tomography (CECT) at initial diagnosis - before any active local or systemic treatment- were exported to our commercially available contouring software (Velocity AI v3.0.1). The volumes of interest (VOIs), including the gross primary tumor volumes (GTVp), were manually segmented by a radiation oncologist in a 3D fashion, then inspected by a second radiation oncologist. The generated VOIs and CT images were exported in the format of DICOM and DICOM-RTSTRUCT to be used for radiomics features extraction.

Radiomics analysis

The primary tumor volumes (GTVp) were contoured based on the ICRU 62/83 definition [12]. Radiomics analysis was performed by the use of the freely available open-source software "Imaging Biomarker Explorer" (IBEX), which was developed by the University of Texas MD Anderson Cancer Center and utilized the MATLAB platform (MathWorks Inc, Natick, VA). The CT images in the format of DICOM and the GTVp contours in the format DICOMRTSTRUCT were imported into IBEX. We extracted features that represent intensity, shape, and texture of a tumor. The categorization of these features was ranked as first, second, and higher texture features based on the applied method from pixel to pixel [13]. More than 3,800 radiomic features were considered in this analysis.

From these radiomic features, we removed those with zero variance and those with a correlation above 99%. Previous studies have identified tumor volume and intensity as relevant features for local control [2]. To further reduce redundancy, we also removed any radiomic features that were highly correlated (>80%) to the features: F25.ShapeVolume and F29.IntensityDirectGlobalMean. Ultimately these resulted in a remaining 301 radiomic features that were used for the proximity computation.

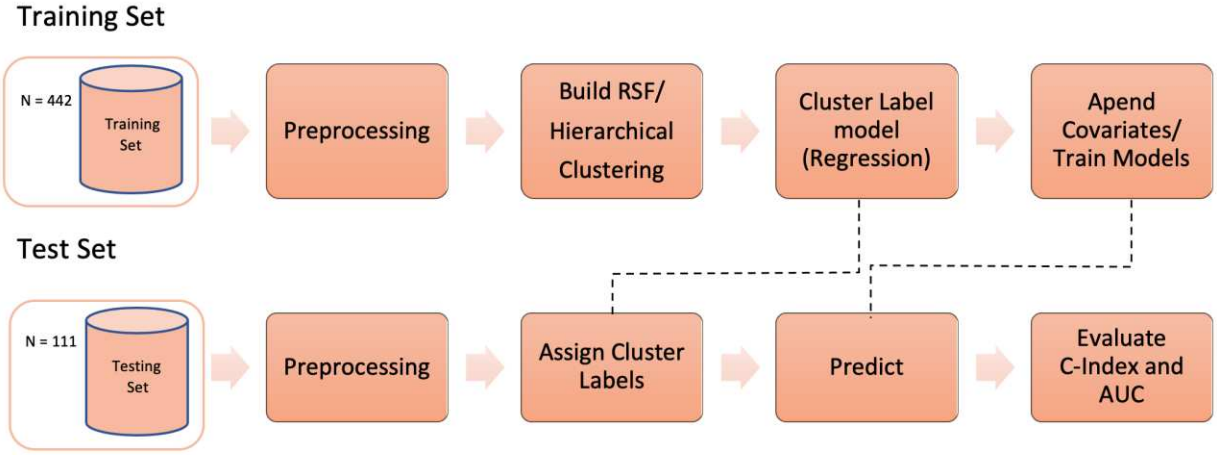
Data preprocessing

As clinical data we consider age (continuous), HPV status (Positive/Negative), Smoking status (Current/Former/Never), T-category 2 groups (T1-2/T3-4), N-category 2 groups (N0-1/N2-3), Therapeutic Combination (CC, IC+CC, IC+Radiation Alone, Radiation Alone), and AJCC staging (8th edition).

Unknown or missing values were imputed using Multivariate Imputation by Chained Equations (MICE) [43].

Methodological Development. Figure 1 shows the overall processing pipeline, including the procedures for feature extraction, evaluation, and cluster explanations. 80% of the sample was used in the training set, and 20% of samples in the test set.

Figure 1. ***Processing pipeline overview.*** *The data is split into disjoint training and validation (test) sets. Initially the data is preprocessed and then the patients are clustered using Random Survival Forest (RSF) clustering. A regression model is trained using the cluster label as dependent variable and later used to assign test patients into a cluster. The ensemble model is trained using clinical covariates and the cluster labels and evaluated over the test data using the discriminaton metrics C-Index and AUC.*



Using the training samples, we fit a random survival forest with the radiomic features as the possible predictors and the right-censored time-to-recurrence as the outcome, i.e., overall or recurrence free survival. We computed the proximity matrix from the random survival forest's fit, i.e., the proportion of times two subjects fall into the same terminal node. Proximities computed for the training set are based on in-bag proximity, i.e., only considering the patients selected across all bootstrap samples. We decided to use in-bag instead of the default out-of-bag samples, because during clustering we are not using the random survival forests as a predictive model but rather to compute the similarity between two very high-dimensional samples

The proximity matrix can be considered a similarity matrix and converted into a dissimilarity measure by subtracting it from the unit matrix. This dissimilarity matrix is then used for clustering, and the clustering algorithm that we use must consider only distances between points and not their absolute positions. Hierarchical clustering [28] is a greedy approach where clusters are built either by starting with one large cluster and splitting it apart (divisive) or starting with a cluster for each point and then merging them at each step (agglomerative). We used the agglomerative approach along with the proximity matrix in our approach. With the matrix, we take the two most similar subjects and cluster them together. Distance between clusters may be measured several ways, and in this study, we used ward [35], which is calculated with the following equation:

$$\delta(c_1, c_2) = \frac{|c_1||c_2|}{|c_1|+|c_2|} \, ||c_1 - c_2||^2 \qquad (1)$$

δ is variance where the goal is to optimize it by minimizing the change, or the error sum of squares. The final extracted feature is simply the resulting cluster label from hierarchical clustering. Survival curves for subjects in each cluster were estimated using the Kaplan-Meier estimator.

After clustering, validation is done using a holdout test set, where test patients are not part of the original clustering. To assign a cluster label to the test samples, we train a regression model over the most important variables from the RSF using a Multinomial Log-Linear Model (mulitnom) to predict the cluster label. Multinomial regression was used instead of the classic binary logistic regression because we want to allow testing for more than 2 clusters.

To assess the added value of the radiomics clusters to predicting survival outcomes beyond standard clinical and demographic characteristics, we compare the performance of a predictive model using only clinical covariates with the same model including both clinical

covariates and the cluster label. We fit an ensemble model using various regression and machine-learning-based models (Cox Proportional Hazard, Random Survival Forest, Random Forest, Logistic Regression, and Logistic-Elastic Net). The first two models are able to handle right-censored outcomes directly, while the later three require a binary outcome. We consider 5-year survival as the event outcome. Only patients that experienced the event before the 5-year cutoff are considered as positive samples. These models have been adapted to right-censored outcomes using inverse probability of censoring weights [17] and patients without sufficient follow-up time that have not experienced the event have zero weight. These prediction models were combined into an ensemble model using stacking. We generated a stacked regression model using the base models' predictions as features and minimizing the prediction error. We use 5-fold cross-validation over the training set to learn the values for the individual models' coefficients (weights) to create the ensemble model. Using the individual model predictions from when each sample was in the test fold, we learn the coefficients that would minimize the square error of the prediction using the non-negative least squares (NNLS) method based on the Lawson-Hanson algorithm and the dual method of Goldfarb from the Superlearner R package [36].

The performance of the ensemble model was assessed using the hold-out test set. In addition to the model using clinical data only, we compare performance to the models including clinical data and AJCC Staging, and a set of raw radiomic features selected using two different supervised methods: Random Survival Forest [15] and Coxnet [37]. For the Random Survival Forest, we use the top features ranked by variable importance (highest frequencies). We use 1000 trees and a default node size of 2. To account for the randomness of the survival forest, we averaged the results after running ten times. The other feature selection method is a Cox Proportional Hazards Model using Regularization Paths for Generalized Linear Models (glm) via

Coordinate Descent (coxnet) [37]. We use cross-validation over the training dataset to find the optimal value for the regularization coefficient and then use it to train the coxnet over the entire dataset and select the features with non-zero coefficients from the model. We use the term COX to represent these features. Two metrics of discrimination are used to evaluate the predictions for all the models: the area under the receiver operating curve (AUC) [44] to predict 5-year survival and Harrel's C-index [45].

Results

Table 1 summarizes the clinical and demographic characteristics of the 533 patients who met the inclusion criteria for this study. The split of training (442) and testing (111) is shown. The cohort was predominately male (~87% for both sets) and the median age was 58 and 56 for training and testing, respectively. Over half of the cohort (>60% for both sets) was HPV positive. ~20% of patients died during follow-up and ~18% experienced a relapse.

Table 1. ***Data demographics****. The table shows the demographics for the clinical covariates used in this study. The dataset (533 patients) was randomly split into training and testing disjoint sets using a 80-20 split. As expected, the same distributions can be observed for the train (442 patients) and test (111 patients) datasets. Within the cells in the table, the reported number is either: count (frequency %) for categorical/discrete covariates, or median (25^{th} – 75^{th} percentiles) for continuous covariates.*

Covariates		
Name	**Train (442)**	**Test (111)**
Gender		
Male	388 (87.8%)	97 (87.4%)
Female	54 (12.2%)	14 (12.6%)
Age at Diagnosis (years)	58.2 (52.5-65.8)	56.6 (52.5-65.8)
T Category		
T1/T2	277 (62.7%)	69 (62.2%)
T3/T4	165 (37.3%)	42 (37.8%)

N Category		
N0/N1	226 (51.1%)	59 (53.2%)
N2/N3	216 (48.8%)	52 (46.8%)
AJCC Stage (8th Edition)		
I	153 (34.6%)	42 (37.8%)
II	82 (18.6%)	17 (15.3%)
III	57 (12.9%)	9 (8.1%)
IV	150 (33.9%)	43 (38.8%)
Smoking Status		
Former	158 (35.8%)	44 (39.7%)
Current	92 (20.8%)	26 (23.4%)
Never	192 (43.4%)	41 (36.9%)
Therapeutic Combination		
CC	228 (51.6%)	68 (61.3%)
IC+CC	119 (26.9%)	26 (23.4%)
IC+Radiation Alone	44 (10.0%)	10 (9.0%)
Radiation Alone	41 (11.5%)	7 (6.3%)
HPV Status		
Positive	288 (65.2%)	67 (60.4%)
Negative	154 (34.8%)	44 (39.6%)
Response		
Vital Status (at end of follow-up)		
Alive	355 (80.3%)	89 (80.2%)
Deceased	87 (19.7%)	22 (19.8%)
Survival Time in months	65.4 (45.9-98.7)	75.3 (48.3-98.1)
Relapse Free Survival		
Alive	363 (82.1%)	91 (82.0%)
Deceased	79 (17.9%)	20 (18.0%)
Survival Time in months	61.0 (40.6-96.4)	69.4 (39.3-94.8)

The Random Survival Forest (RSF) was built over the training data and log-rank was used as the splitting rule, with a minimum node size of 5 as previously used to predict Parkinson's disease with radiomic data [27]. The number of trees per forest was set to 1000. The co-occurrence matrix was extracted from the RSF and hierarchical clustering was used to identify 2-4 groups. Overall, the clusters were more balanced for OS than for RFS. For 2 clusters, the split was roughly 50-50% for OS and 70-30% for RFS.

Figure 2 and 3 shows the Kaplan-Meier survival curves for the training and test patients stratified by the proposed cluster labels for OS and RFS, respectively. These results show that the similarity and the subsequent hierarchical clustering are sensible means to capturing radiomic feature differences. For both clustering outcomes, there is a visible separation between the groups.

Figure 2. ***Kaplan-Meier (KM) Curves for Overall Survival (OS).*** *The figure shows the KM curves for OS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves are significantly different which indicates that the proposed clustering is effective in identifying a risk stratification and can be effectively used as a predictive covariate.*

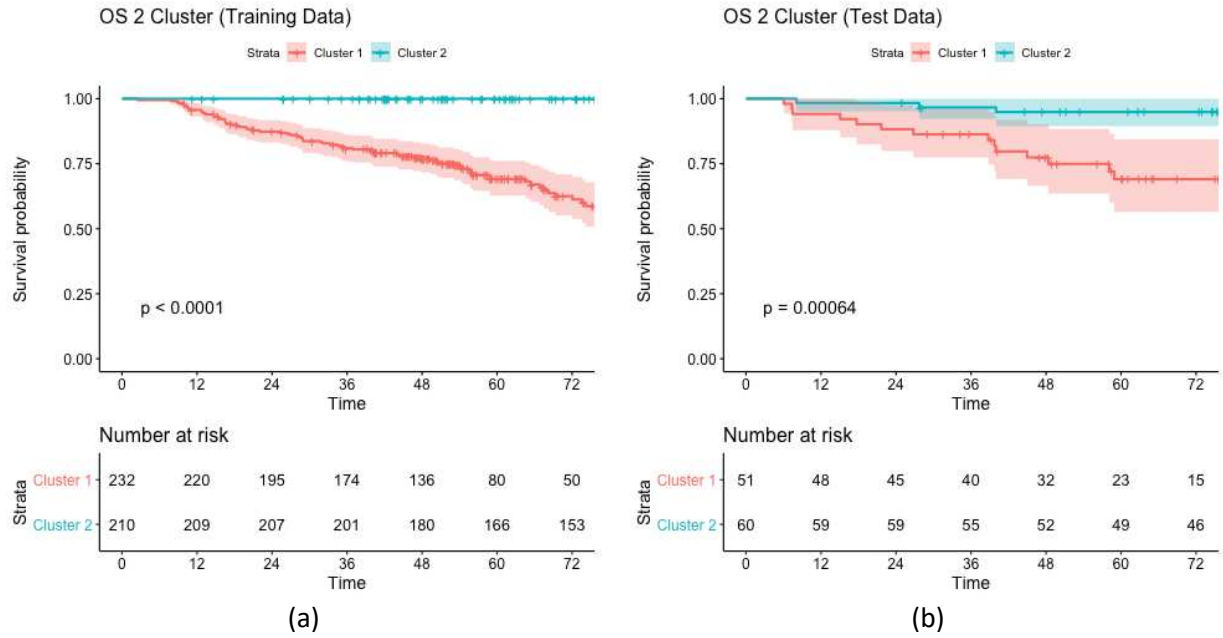
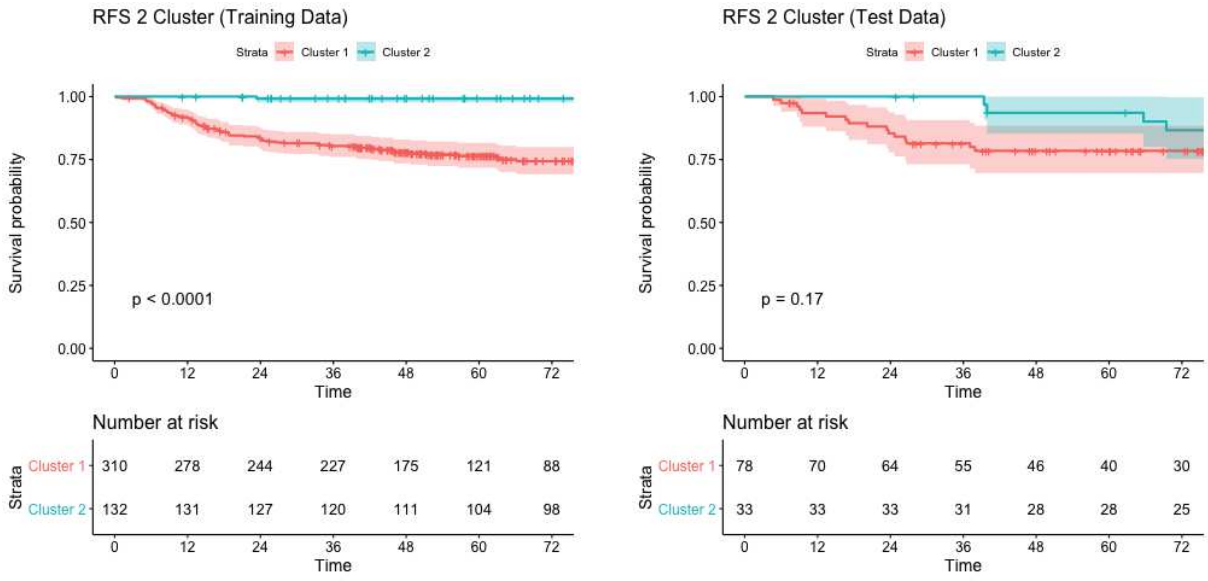


Figure 3. ***Kaplan-Meier (KM) Curves for Recurrence Free Survival (RFS).*** *The figure shows the KM curves for RFS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves show two consistent risk groups which indicates that the proposed clustering can be effectively used as a predictive covariate within a risk prediction model.*



(a) (b)

Figure 2 shows the curves of the OS outcome for 2 clusters. As can be seen, for both training and testing results, the proposed approach is successful in stratifying the patients by their survival risk. The survival curves for the two clusters are significantly different for both training and testing (p value < 0.005).

Figure 3 above shows the curves of the RFS outcome for 2 clusters. As can be seen there is separation between the curves for both the training and testing sets. While the training shows significantly different curves (p-value < 0.0001), the separation between the survival curves for two clusters over the test dataset is not statistically significant (p-value=0.17).

The supervised feature selection algorithms were used for comparison with the cluster label. The top 10 ranked features using variable importance (highest frequency) for the Random Survival Forest were selected for both OS and RFS. The Cox Proportional Hazards Model using Regularization Paths for Generalized Linear Models via Coordinate Descent (coxnet) identified 5 and 8 radiomic features for OS and RFS, respectively. Table 2 lists all the features names used as predictive covariates in the ensemble model.

Table 2. ***Covariates used in the ensemble model.*** *The clinical covariates are used independently of the outcome being evaluated. Since Random Survival Forests (RSF) and Coxnet (COX) can be used as supervised feature selection methods, the radiomic features selected depend on the outcome used. The top 10 covariates from RSF are selected for each outcome. For COX, the features selected depend on the number of non-zero weights learned by the regularization coefficient. COX selected 5 and 8 radiomics features for OS and RFS, respectively. Cluster refers to the cluster label extracted using Random Survival Forest Clustering.*

Name	*Count*	*Covariates*
Clinical	*6*	*Age, HPV status (positive | negative), Smoking Status (never | former | current), T.category ([T1-T2],[T3-T4]), N.category ([N0-N1],[N2-N3]), Therapeutic Combination (RT alone, Concurrent Chemotherapy (CC) , Induction+RT, Induction+CC)*
RSF (OS)	*10*	*F4.GrayLevelRunLengthMatrix25..90ShortRunLowGrayLevelEmpha,F48.GrayLevelCooccurenceMatrix25180.2ClusterProminence,F48.GrayLevelCooccurenceMatrix25270.1Contrast,F48.GrayLevelCooccurenceMatrix25225.7ClusterShade,F29.IntensityDirectLocalRangeMax,F2.GrayLevelCooccurenceMatrix25270.1Contrast,F2.GrayLevelCooccurenceMatrix25.333.4Correlation,F2.GrayLevelCooccurenceMatrix25180.6MaxProbability,F4.GrayLevelRunLengthMatrix25..90RunLengthNonuniformity,F4.GrayLevelRunLengthMatrix25.333ShortRunEmphasis*
RSF (RFS)	*10*	*F48.GrayLevelCooccurenceMatrix25180.2ClusterProminence,F48.GrayLevelCooccurenceMatrix25315.6ClusterProminence,F8.IntensityDirectKurtosis, F9.IntensityDirectSkewness,F11.IntensityDirectKurtosis, F13.IntensityDirectEnergy,F48.GrayLevelCooccurenceMatrix25180.1InverseDiffNorm,F2.GrayLevelCooccurenceMatrix25180.5ClusterProminence,F2.GrayLevelCooccurenceMatrix25180.5ClusterShade,F14.IntensityDirectEnergy*
COX (OS)	*5*	*F25.ShapeVolume, F29.IntensityDirectLocalRangeMax,F4.GrayLevelRunLengthMatrix25..90RunLengthNonuniformity,F6.IntensityDirectSkewness,F48.GrayLevelCooccurenceMatrix25225.7AutoCorrelation*
COX (RFS)	*8*	*F5.IntensityDirectGlobalMax, F13.IntensityDirectGlobalMax,F14.IntensityDirectGlobalMax, F25.ShapeVolume,F29.IntensityDirectLocalRangeMax,F4.GrayLevelRunLengthMatrix25..90RunLengthNonuniformity,F4.GrayLevelRunLengthMatrix25..90ShortRunLowGrayLevelEmpha,F48.GrayLevelCooccurenceMatrix25225.7AutoCorrelation*
Cluster	*1*	*Cluster label with 2, 3, or 4 values*

Figure 4. ***Top Radiomic Features identified by the Random Survival Forest (RSF) for Overall Survival (OS).*** *Boxplots of top 9 features selected using the variable importance from the Random Survival Forest (RSF) over the training data and their distribution within the two clusters identified for Overall Survival (OS). The difference in distribution suggests that these variables can be used in a model to assign cluster labels to test patients.*

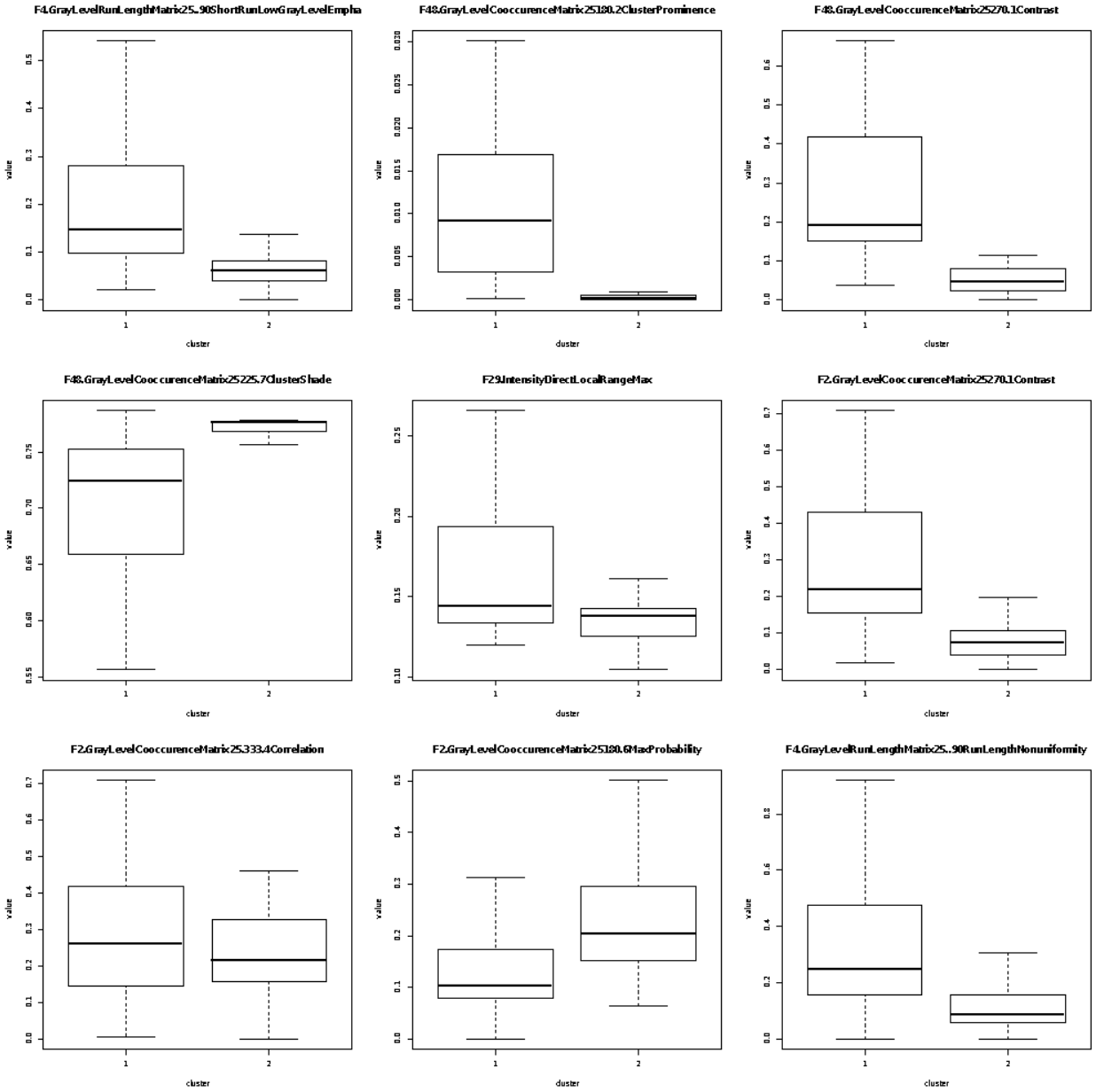


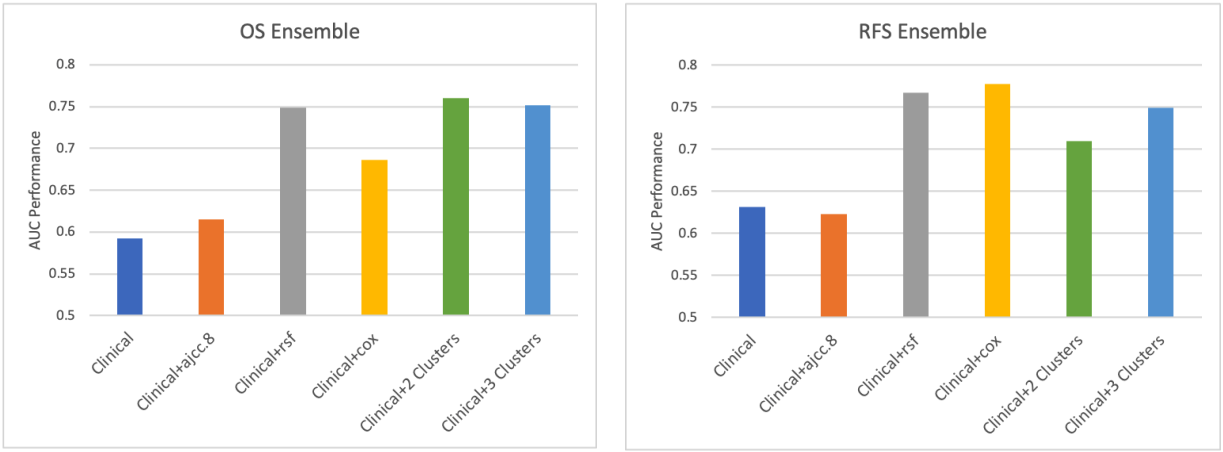
Figure 4 shows the boxplot for the top nine radiomic features for OS within each cluster. From the figure it can be seen that the distribution of these features is different between the two clusters, which makes them good candidates for relevant features to train a cluster assignment

model to label the test samples. A similar result can be seen in the box plots for RFS outcome (See Appendix A, Figure A1).

Figure 5 compares the ensemble AUC performance for the different predictive covariates over the hold-out test dataset for (a) OS and (b) RFS outcomes. The baseline model is denoted as Clinical and is the one trained using the six clinical covariates: Age, HPV status, Smoking status, Therapeutic Combination, T-Category, and N-Category (8th edition). The Clinical+AJCC.8 model is the baseline model when AJCC Staging (8th edition) is added as a predictive covariate. Clinical+rsf denotes the baseline model when the 10 RSF selected radiomic features are also included in the model. Clinical+cox represents the ensemble where the coxnet selected features have been added to the model. Finally, Clinical+N Clusters is the ensemble model when the radiomic cluster (with N groups) has been added as a predictive covariate. When only clinical covariates are used, AUC over the test data is .58 and .63 for OS and RFS, respectively. Compared to clinical only, models that incorporated the cluster label (Clinical+2 Clusters and Clinical+3 Clusters) as a covariate led to substantial improvement in discrimination. The inclusion of three radiomics derived clusters improves performance by over 17% and 12% (AUC=.75) for OS and RFS, respectively. Compared to models which incorporated selected radiomic features directly (+RSF and +COX), discrimination performance was comparable (+1% for OS and within 2-3% for RFS).

Figure 5. ***Ensemble model performance over test data.*** *The ensemble model discrimination was evaluated using the AUC metric over the test data for two survival outcomes: (a) Overall Survival (OS) and (b) Recurrence Free Survival (RFS). Comparison is done between a Clinical baseline model using six clinical covariates: age, hpv status, smoking status, T-category, N-category, and therapeutic combination and the models including additional model covariates: AJCC staging (Clinical+ajcc.8), selected radiomic features (Clinical+rsf/+cox), and the proposed cluster labels (Clinical+N Clusters). In all cases, the inclusion of the cluster labels outperforms Clinical and Clinical+ajcc.8 models. For OS, the best performing*

model is the Clinical+2 Clusters model with AUC:.76. For RFS, the Clinical+3 Clusters reaches AUC:0.75, comparable to the Clinical+rsf/+cox models (AUC:.78) while being a considerably more parsimonious model.



(a) (b)

Table 3. ***Ensemble discrimination performance over training and testing data.*** *Comparison of ensemble performance over Train and Test data using C-Index and AUC for both OS and RFS outcomes. Each row in the table corresponds to the ensemble model using different covariates. The Clinical baseline is the model where only clinical covariates are included. The subsequent rows include additional covariates into the baseline model: AJCC staging (8th edition) (+ajcc.8), selected radiomic features (+rsf, +cox), and the proposed cluster labels (Clinical+N Clusters). The best test results are highlighted in bold. The best test results for OS are obtained by the Clinical+2 Cluster model (C-Index: .71, AUC: .76) while the best test results for RFS are obtained by Clinical+cox (C-Index: .71, AUC: .78). Clinical+cox model has a total of 14 covariates including 9 continuous variables (age + 8 radiomic features) while Clinical+3 Clusters (C-Index: .69, AUC: .75) only has a total of 7 covariates with only age being continuous.*

Covariates Used in Model	Overall Survival (OS)				Recurrence Free Survival (RFS)			
	C-Index		AUC		C-Index		AUC	
	Train	Test	Train	Test	Train	Test	Train	Test
Clinical	.66	.58	.66	.59	.61	.61	.68	.63
Clinical+ajcc.8	.65	.61	.65	.62	.62	.61	.68	.62
Clinical+rsf	.75	.69	.79	.75	.66	.69	.73	.77
Clinical+cox	.73	.65	.75	.69	.64	**.71**	.71	**.78**
Clinical+2 Clusters	.77	**.71**	.82	**.76**	.71	.65	.80	.71
Clinical+3 Clusters	.78	.69	.83	.75	.79	.69	.89	.75
Clinical+4 Clusters	.90	.64	.93	.67	.95	.54	.98	.55

Table 3 shows the ensemble performance using C-Index and AUC over training and test data for both outcomes. It is worth noting that while Clinical+2 Clusters and Clinical+3 Clusters results are comparable and close to each other, Clinical+4 Clusters seem to be overfitting the data, with the highest training performance and the lowest testing performance.

Discussion

The proposed method for clustering the high-dimensional radiomic features using hierarchical clustering over the co-occurrence matrix extracted from a Random Survival Forest (RSF) model is a sensible way to summarize the radiomic features into a single covariate. The hierarchical clustering method is robust and generates informative clusters across the different outcomes. The use of a regression model over the most important (frequent) variables selected from the RSF to assign a cluster label offers a simple yet effective way to label the previously unseen test samples. For OS, the Kaplan-Meier survival curves show statistically significant separation between the curves for both training and testing (p-value < 0.01, Figure 2). For RFS, even when the test curves follow the same risk stratification as the training curves, the separation between the curves is not statistically significant (p-value=0.17, Figure 3). A possible explanation for this performance for RFS may be due to the fact that RFS is a combined outcome and only the radiomic features from the primary tumor were considered for clustering. Nevertheless, as can be seen in the model evaluation, the addition of the RFS clusters to other predictive clinical covariates including N-staging, HPV status, and Therapeutic combination improves model performance for both training and testing. Prior work has also effectively leveraged clustering to improve outcome prediction for OPC patients [39-41, 49], however, none of these works have attempted to use the entire set of radiomic features or Random Survival Forest learning as we have done in this work.

Including the proposed cluster labels as a predictive covariate considerably improves model discrimination for survival outcomes when compared to the same model using clinical data only. Moreover, the performance for models including the radiomic clusters is comparable

to the models including radiomic features selected using supervised algorithms (RSF and Coxnet). Several studies on head and neck cancer data have identified radiomic signatures using machine learning approaches to improve different survival outcomes [46-48]. While these algorithms select a small number of radiomic features (up to 10 continuous variables in our experiments), the number of radiomic features is still sometimes larger than the clinical covariates included in the model. In contrast, our cluster label approach yields a more parsimonious model that uses one categorical variable with only 2 or 3 values. Having a smaller subset of features to represent the radiomics is especially useful when there is a small to moderate number of samples, the event rate is low (e.g., under 20% in our case), and few clinical covariates are added into the model (5 in our case).

The results for OS show that a single covariate to represent the high-dimensional radiomic features can be more predictive than a handful of selected features. The reason is that the cluster label offers a better generalization by reducing the noise and sparsity of the data. Moreover, with the proposed feature extraction method, we are able to easily analyze and stratify the populations based on their cluster labels. An additional benefit of random survival forest clustering is that no feature scaling is required because random forest algorithms are not affected by monotonic transformations. Since the output is a categorical cluster label no scaling is required when training any models either. With feature selection, scaling may be required during selection if methods besides random forest are used either during selection or model training. When a very low-dimensional explanation of radiomic data is required, we recommend the use of feature extraction via random forest clustering, and furthermore, we recommend hierarchical clustering to obtain reasonably balanced clusters.

The main limitations of this work derive from the small sample size and the large number of right-censored samples. Because of these factors, we are not able to evaluate the proposed feature extraction with a large number of clusters or conclude anything about the optimal cluster size. Instead, the number of clusters was varied from 2, because it is the fewest number of clusters, up to 4 because of the categorization of the primary tumor, T category, which typically has 4 categories and because our radiomic feature set is based on the primary tumor. However, while the results were comparable between 2 and 3 clusters, 4 clusters suffered from overfitting in our experiments. Some radiomic clustering studies have used techniques to determine an optimal number of clusters using Principal Component Analysis (PCA) and cluster validation [35] or consensus clustering [6]. In Zdilar et al. [42], different transformations for right-censored survival outcomes are considered, one of which consists of using the Martingale residuals obtained from a Cox Proportional Hazards model. The Martingale residual can be considered as a continuous outcome. As a potential future work, we could use the Martingale residuals as an outcome and apply the same methodology described in this work using Random Regression Forests [34] to generate a patient-to-patient similarity matrix for clustering.

In conclusion, feature extraction via random survival forest clustering greatly reduces the radiomic feature space and compares well to feature selection in predictive performance for survival outcomes. Feature extraction in this way can be particularly beneficial when it is desirable to have a very concise representation of the radiomic feature space such as when the number of features is low, or the number of clinical features is already high and the number of samples is moderate to low.

Data Availability

The datasets analyzed during the current study are available from Scientific Data [38] and TCGA.

Ethics declarations

Competing interests

The authors declare no competing interests.

Author Information

Affiliations

University of Iowa, Department of Electrical and Computer Engineering, Iowa City, 52242, USA

Harsh, Patel & Guadalupe Canahuate

University of Minnesota, Division of Biostatistics, Minneapolis, 55455, USA

David M. Vock

University of Illinois at Chicago, Department of Department of Computer Science, Chicago, 60607, USA

G. Elisabeta Marai

MD Anderson Cancer Center, Department of Radiation Oncology, Houston, 77030, USA

Clifton Fuller & Abdallah S. R. Mohamed

Contribution

Specific additional individual cooperative effort contributions to study/manuscript design/execution/interpretation, in addition to all criteria above are listed as follows: H.P. Coded and conducted the experiments, undertook supervised analysis and interpretation of the data, and oversought the manuscript execution. A.S.R.M., Undertook clinical data collection; executed and quality assured data collection workflow; and participated in data analysis, interpretation, and manuscript drafting and final editing. D.V., G.E.M. and C.D.F. Provided direct analysis, statistical and mathematical modeling, and data interpretation expertise. G.C.- Corresponding author; conceived, coordinated, and directed all study activities, project integrity, manuscript

content and editorial oversight and correspondence; direct oversight of trainee personnel. All listed co-authors performed the following: 1. Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work. 2. Drafting the work or revising it critically for important intellectual content. 3. Final approval of the version to be published. 4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Corresponding author

Correspondence to guadalupe-canahuate@uiowa.edu

References

[1] K. M. Panth, R. T. Leijenaar, S. Carvalho, N. G. Lieuwes, A. Yaromina, L. Dubois, P. Lambin, Is there a causal relationship between genetic changes and radiomics-based image features? an in vivo preclinical experiment with doxycycline inducible gadd34 tumor cells, Radiotherapy and Oncology 116 (3) (2015) 462–466.

[2] M. A. C. C. Head, N. Q. I. W. Group, et al., Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients, Scientific reports 8.

[3] H. Aerts, E. R. Velazquez, R. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Mon- shouwer, B. Haibe-Kains, D. Rietveld, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun 5 (4006) (2014) 1–8.

[4] Y. Huang, Z. Liu, L. He, X. Chen, D. Pan, Z. Ma, C. Liang, J. Tian, C. Liang, Radiomics signature: A potential biomarker for the prediction of disease-free survival in early-stage (i or ii) nonsmall cell lung cancer, Radiology 281 (3) (2016) 947–957.

[5] A. J. Wong, A. Kanwar, A. S. Mohamed, C. D. Fuller, Radiomics in head and neck cancer: from exploration to application, Translational Cancer Research 5 (4) (2016) 371–382. [6] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, H. J. Aerts, Radiomic machine- learning classifiers for prognostic biomarkers of head and neck cancer, Frontiers in oncology 5.

[6] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Z¨ophel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, et al., A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling, Scientific Reports 7 (1) (2017) 13206.

[7] M. Valli`eres, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, K. Sultanem, et al., Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, arXiv preprint arXiv:1703.08516.

[8] L. Lu, R. C. Ehmke, L. H. Schwartz, B. Zhao, Assessing agreement between radiomic features computed for multiple ct imaging settings, PLoS One 11 (12) (2016) e0166550.

[9] Q. Zhang, Y. Xiao, J. Suo, J. Shi, J. Yu, Y. Guo, Y. Wang, H. Zheng, Sonoelastomics for breast tumor classification: a radiomics approach with clustering-based feature selection on sonoelastography, Ultrasound in Medicine and Biology 43 (5) (2017) 1058–1069.

[10] H. Li, M. Galperin-Aizenberg, D. Pryma, C. Simone, Y. Fan, Unsupervised machine learning of radiomic features for predicting treatment response and survival of early-stage nonsmall cell lung cancer patients treated with stereotactic body radiation therapy, International Journal of Radiation Oncology* Biology* Physics 99 (2) (2017) S34.

[11] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider, F. Khalvati, Radiomics-based prognosis analysis for non-small cell lung cancer, Scientific reports 7 (2017) 46349.

[12] P. Royston, D. G. Altman, External validation of a cox prognostic model: principles and methods, BMC medical research methodology 13 (1) (2013) 33.

[13] T. M. Therneau, P. M. Grambsch, T. R. Fleming, Martingale-based residuals for survival models, Biometrika 77 (1) (1990) 147–160.

[14] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests, The annals of applied statistics (2008) 841–860.

[15] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[16] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, P. J. OConnor, Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting, Journal of biomedical informatics 61 (2016) 119–131.

[17] T. Shi, D. Seligson, A. S. Belldegrun, A. Palotie, S. Horvath, Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Modern Pathology 18 (4) (2005) 547.

[18] E. Allen, S. Horvath, F. Tong, P. Kraft, E. Spiteri, A. D. Riggs, Y. Marahrens, High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes, Proceedings of the National Academy of Sciences 100 (17) (2003) 9940–9945.

[19] L. Breiman, A. Cutler, Random forests manual v4, in: Technical report, UC Berkel, 2003.

[20] R. M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (5) (1979) 786–804.

[21] M. H. Bharati, J. J. Liu, J. F. MacGregor, Image texture analysis: methods and comparisons, Chemo- metrics and intelligent laboratory systems 72 (1) (2004) 57–71.

[22] D. Mackin, L. Court, C. Ng, J. Yang, L. Zhang, X. Fave, Su-f-r-09: Homogenization of ct images for radiomics studies: It's like butter (worth), Medical physics 43 (6) (2016) 3374–3375.

[23] X. Fave, L. Zhang, J. Yang, D. Mackin, P. Balter, D. Gomez, D. Followill, A. K. Jones, F. Stingo, et al., Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer, Translational Cancer Research 5 (4) (2016) 349–363.

[24] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, K. Miles, Tumour heterogeneity in oesophageal cancer assessed by ct texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival, Clinical radiology 67 (2) (2012) 157–164.

[25] F. Davnall, C. S. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook, V. Goh, Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?, Insights into imaging 3 (6) (2012) 573–589.

[26] A. Rahmim, P. Huang, N. Shenkov, S. Fotouhi, E. Davoodi-Bojd, L. Lu, Z. Mari, H. Soltanian-Zadeh, V. Sossi, Improved prediction of outcome in parkinson's disease using radiomics analysis of longitudinal dat spect images, NeuroImage: Clinical 16 (2017) 539–544.

[27] J. Franklin, The elements of statistical learning: data mining, inference and prediction, The Mathe- matical Intelligencer 27 (2) (2005) 83–85.

[28] S. C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (3) (1967) 241–254.

[29] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of bioinformatics and computational biology 3 (02) (2005) 185–205.

[30] N. De Jay, S. Papillon-Cavanagh, C. Olsen, G. Bontempi, B. Haibe-Kains, mrmre: an r package for parallelized mrmr ensemble feature selection, Submitted (2012) .

[31] V. Bewick, L. Cheek, J. Ball, Statistics review 12: survival analysis, Critical care 8 (5) (2004) 389.

[32] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (5) (2002) 429–449.

[33] J. H. Friedman, On bias, variance, 0/1loss, and the curse-of-dimensionality, Data mining and knowledge discovery 1 (1) (1997) 55–77.

[34] Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. Biometrika, 77(1), 147–160. https://doi.org/10.1093/biomet/77.1.147

[35] Murtagh, F., & Legendre, P. (2011). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *ArXiv, abs/1111.6285*.

[36] van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (16 Sep. 2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1). https://doi.org/https://doi.org/10.2202/1544-6115.1309

[37] Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. Statistics in Medicine, 24(11), 1713–1723. https://doi.org/https://doi.org/10.1002/sim.2059

[38] Elhalawani, H. et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci. data 4, 170077 (2017).

[39] Wentzel, A., Hanula, P., van Dijk, L.V., Elgohari, B., Mohamed, A.S., Cardenas, C.E., Fuller, C.D., Vock, D.M., Canahuate, G. and Marai, G.E., 2020. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. Radiotherapy and Oncology.

[40] Luciani, T., Wentzel, A., Elgohari, B., Elhalawani, H., Mohamed, A., Canahuate, G., Vock, D.M., Fuller, C.D. and Marai, G.E., 2020. A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine. Journal of Biomedical Informatics, 5, p.100067.

[41] Tosado, J., Zdilar, L., Elhalawani, H. et al. Clustering of Largely Right-Censored Oropharyngeal Head and Neck Cancer Patients for Discriminative Groupings to Improve Outcome Prediction. Sci Rep 10, 3811 (2020). https://doi.org/10.1038/s41598-020-60140-0

[42] Zdilar, Luka, et al. “Evaluating the Effect of Right-Censored End Point Transformation for Radiomic Feature Selection of Data From Patients With Oropharyngeal Cancer.” JCO Clinical Cancer Informatics, no. 2, Dec. 2018, pp. 1–19. DOI.org (Crossref), doi:10.1200/CCI.18.00052.

[43] Buuren, Stef van, and Karin Groothuis-Oudshoorn. “Mice : Multivariate Imputation by Chained Equations in R.” Journal of Statistical Software, vol. 45, no. 3, 2011. DOI.org (Crossref), doi:10.18637/jss.v045.i03.

[44] Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users’ Guides to the Medical Literature. JAMA. 2017;318(14):1377–1384. doi:10.1001/jama.2017.12126

[45] Harrell , Frank E. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing, 2015.

[46] MDACC Head, Neck Quantitative Imaging Working Group. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. Scientific Reports 8

[47] Elhalawani, H., Lin, T. A., Volpe, S., Mohamed, A. S., White, A. L., Zafereo, J., ... & Aymard, J. M. (2018). Machine learning applications in head and neck radiation oncology: lessons from open-source radiomics challenges. Frontiers in oncology, 8, 294.

[48] Marai, G. E., Ma, C., Burks, A. T., Pellolio, F., Canahuate, G., Vock, D. M., ... & Fuller, C. D. (2018). Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Trans. on Vis. and Comp. Graphics, 25(4), 1732-1745.

[49] Wentzel, A., Hanula, P., Luciani, T., Elgohari, B., Elhalawani, H., Canahuate, G., ... & Marai, G. E. (2019). Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration. IEEE Trans. on Vis. and Comp. Graphics, 26(1), 949-959.

[50] Sheu, T., Vock, D. M., Mohamed, A. S., Gross, N., Mulcahy, C., Zafereo, M., ... & Lewin, J. S. (2017). Conditional survival analysis of patients with locally advanced laryngeal cancer: construction of a dynamic risk model and clinical nomogram. Scientific reports, 7, 43928.

Figures Legend

Figure 1. ***Processing pipeline overview.*** *The data is split into disjoint training and validation (test) sets. Initially the data is preprocessed and then the patients are clustered using Random Survival Forest (RSF) clustering. A regression model is trained using the cluster label as dependent variable and later used to assign test patients into a cluster. The ensemble model is trained using clinical covariates and the cluster labels and evaluated over the test data using the discriminaton metrics C-Index and AUC.*

Table 1. ***Data demographics****. The table shows the demographics for the clinical covariates used in this study. The dataset (533 patients) was randomly split into training and testing disjoint sets using a 80-20 split. As expected, the same distributions can be observed for the train (442 patients) and test (111 patients) datasets. Within the cells in the table, the reported number is either: count (frequency %) for categorical/discrete covariates, or median (25^{th} – 75^{th} percentiles) for continuous covariates.*

Figure 2. ***Kaplan-Meier (KM) Curves for Overall Survival (OS).*** *The figure shows the KM curves for OS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves are significantly different which indicates that the proposed clustering is effective in identifying a risk stratification and can be effectively used as a predictive covariate.*

Figure 3. ***Kaplan-Meier (KM) Curves for Recurrence Free Survival (RFS).*** *The figure shows the KM curves for RFS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves show two consistent risk groups which indicates that the proposed clustering can be effectively used as a predictive covariate within a risk prediction model.*

Table 2. ***Covariates used in the ensemble model.*** *The clinical covariates are used independently of the outcome being evaluated. Since Random Survival Forests (RSF) and Coxnet (COX) can be used as supervised feature selection methods, the radiomic features selected depend on the outcome used. The top 10 covariates from RSF are selected for each outcome. For COX, the features selected depend on the number of non-zero weights learned by the regularization coefficient. COX selected 5 and 8 radiomics features for OS and RFS, respectively. Cluster refers to the cluster label extracted using Random Survival Forest Clustering.*

Figure 4. ***Top Radiomic Features identified by the Random Survival Forest (RSF) for Overall Survival (OS).*** *Boxplots of top 9 features selected using the variable importance from the Random Survival Forest (RSF) over the training data and their distribution within the two clusters identified for Overall Survival (OS). The difference in distribution suggests that these variables can be used in a model to assign cluster labels to test patients.*

Figure 5. ***Ensemble model performance over test data.*** *The ensemble model discrimination was evaluated using the AUC metric over the test data for two survival outcomes: (a) Overall Survival (OS) and (b) Recurrence Free Survival (RFS). Comparison is done between a Clinical baseline model using six clinical covariates: age, hpv status, smoking status, T-category, N-category, and therapeutic combination and the models including additional model covariates: AJCC staging (Clinical+ajcc.8), selected radiomic features (Clinical+rsf/+cox), and the proposed cluster labels (Clinical+N Clusters). In all cases, the inclusion of the cluster labels outperforms Clinical and Clinical+ajcc.8 models. For OS, the best performing model is the Clinical+2 Clusters model with AUC:.76. For RFS, the Clinical+3 Clusters reaches AUC:0.75, comparable to the Clinical+rsf/+cox models (AUC:.78) while being a considerably more parsimonious model.*

Table 3. ***Ensemble discrimination performance over training and testing data.*** *Comparison of ensemble performance over Train and Test data using C-Index and AUC for both OS and RFS outcomes. Each row in the table corresponds to the ensemble model using different covariates. The Clinical baseline is the model where only clinical covariates are included. The subsequent rows include additional covariates into the baseline model: AJCC staging (8th edition) (+ajcc.8), selected radiomic features (+rsf, +cox), and the proposed cluster labels (Clinical+N Clusters). The best test results are highlighted in bold. The best test results for OS are obtained by the Clinical+2 Cluster model (C-Index: .71, AUC: .76) while the best test results for RFS are obtained by Clinical+cox (C-Index: .71, AUC: .78). Clinical+cox model has a total of 14 covariates including 9 continuous variables (age + 8 radiomic features) while Clinical+3 Clusters (C-Index: .69, AUC: .75) only has a total of 7 covariates with only age being continuous.*

Acknowledgements

We would like to acknowledge Joel Tosado and Luka Zdilar, who as graduate students at the University of Iowa were involved in the initial stages of this work and participated in the discussions of the approach. We would also like to acknowledge the numerous MD Anderson Cancer Center researchers who contributed to produce and curate the data used in this paper: Hesham Elhalawani, Baher Elgohari, Carly Tiras, Austin Miller, Aasheesh Kanwar, Aubrey White, James Zafereo, Andrew Wong, Joel Berends, Shady Abohashem, Bowman Williams, Jeremy M. Aymard, Subha Perni, Jay Messer, and Ben Warren.

Figures

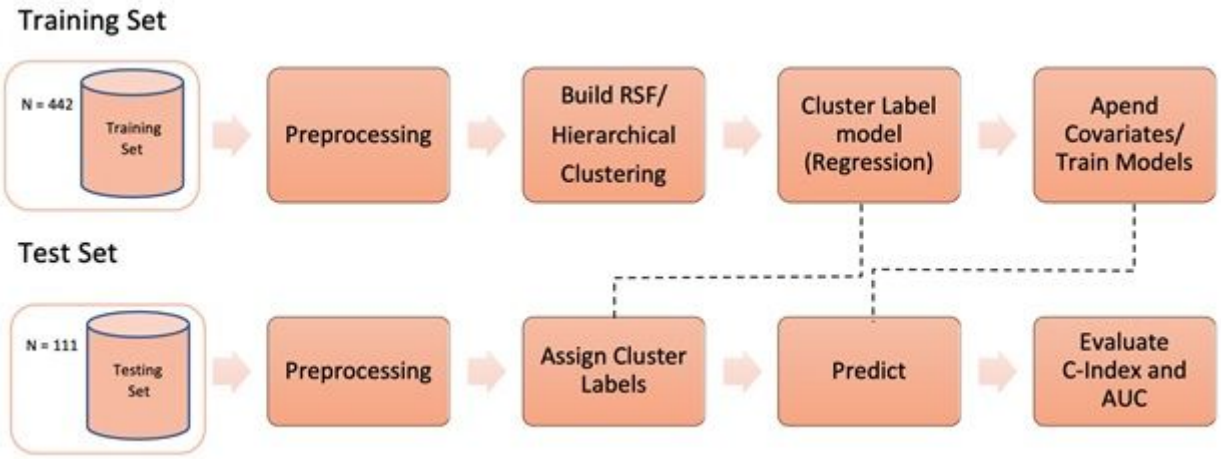


Figure 1

Processing pipeline overview. The data is split into disjoint training and validation (test) sets. Initially the data is preprocessed and then the patients are clustered using Random Survival Forest (RSF) clustering. A regression model is trained using the cluster label as dependent variable and later used to assign test patients into a cluster. The ensemble model is trained using clinical covariates and the cluster labels and evaluated over the test data using the discriminaton metrics C-Index and AUC.

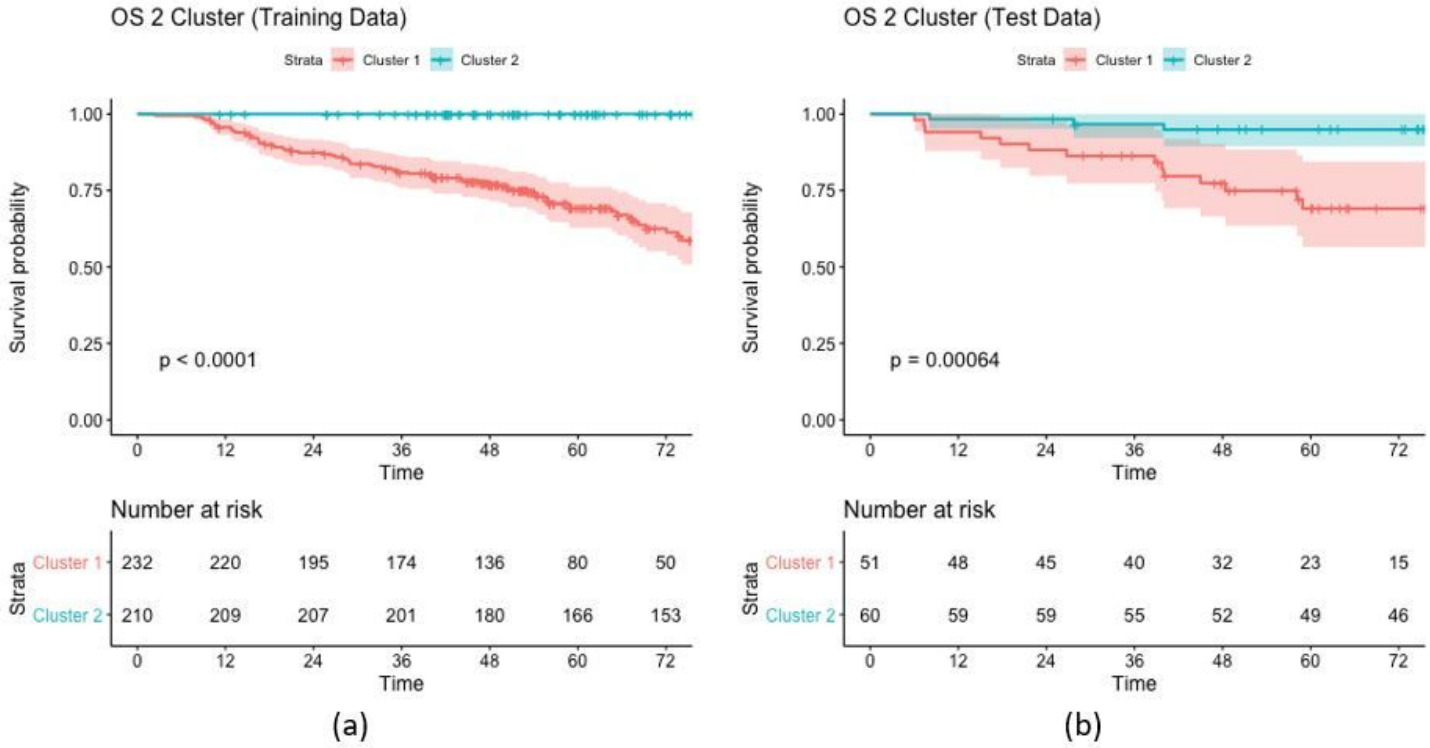


Figure 2

Kaplan-Meier (KM) Curves for Overall Survival (OS). The figure shows the KM curves for OS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves are significantly different which indicates that the proposed clustering is effective in identifying a risk stratification and can be effectively used as a predictive covariate.

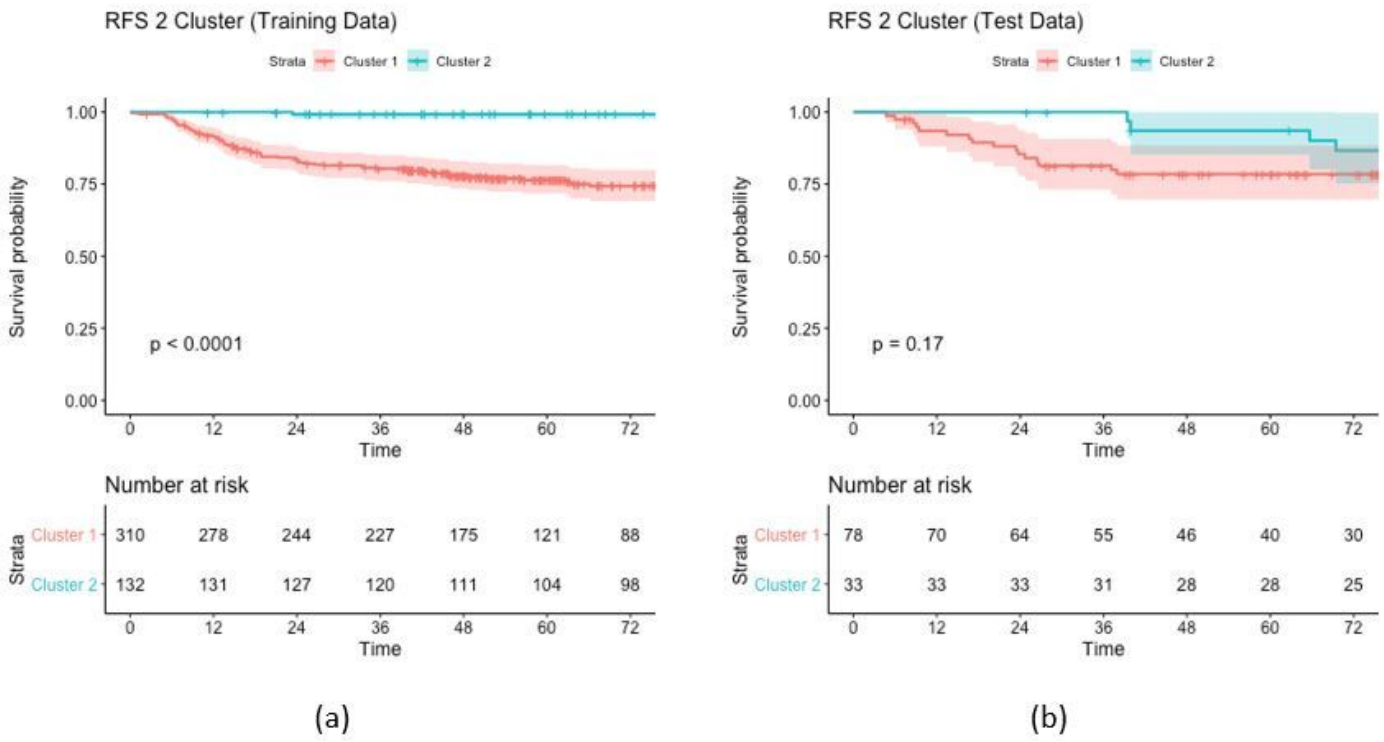


Figure 3

Kaplan-Meier (KM) Curves for Recurrence Free Survival (RFS). The figure shows the KM curves for RFS outcome stratified by the cluster label over (a) training and (b) test data. For the training, the patients were grouped using Hierarchical Clustering over the co-occurrence matrix from the Random Survival Forest. For the testing, the patients were assigned to a cluster by applying the regression model trained for predicting the cluster labels using the top 10 radiomic features identified by the random survival forest. For both training and testing, the KM curves show two consistent risk groups which indicates that the proposed clustering can be effectively used as a predictive covariate within a risk prediction model.

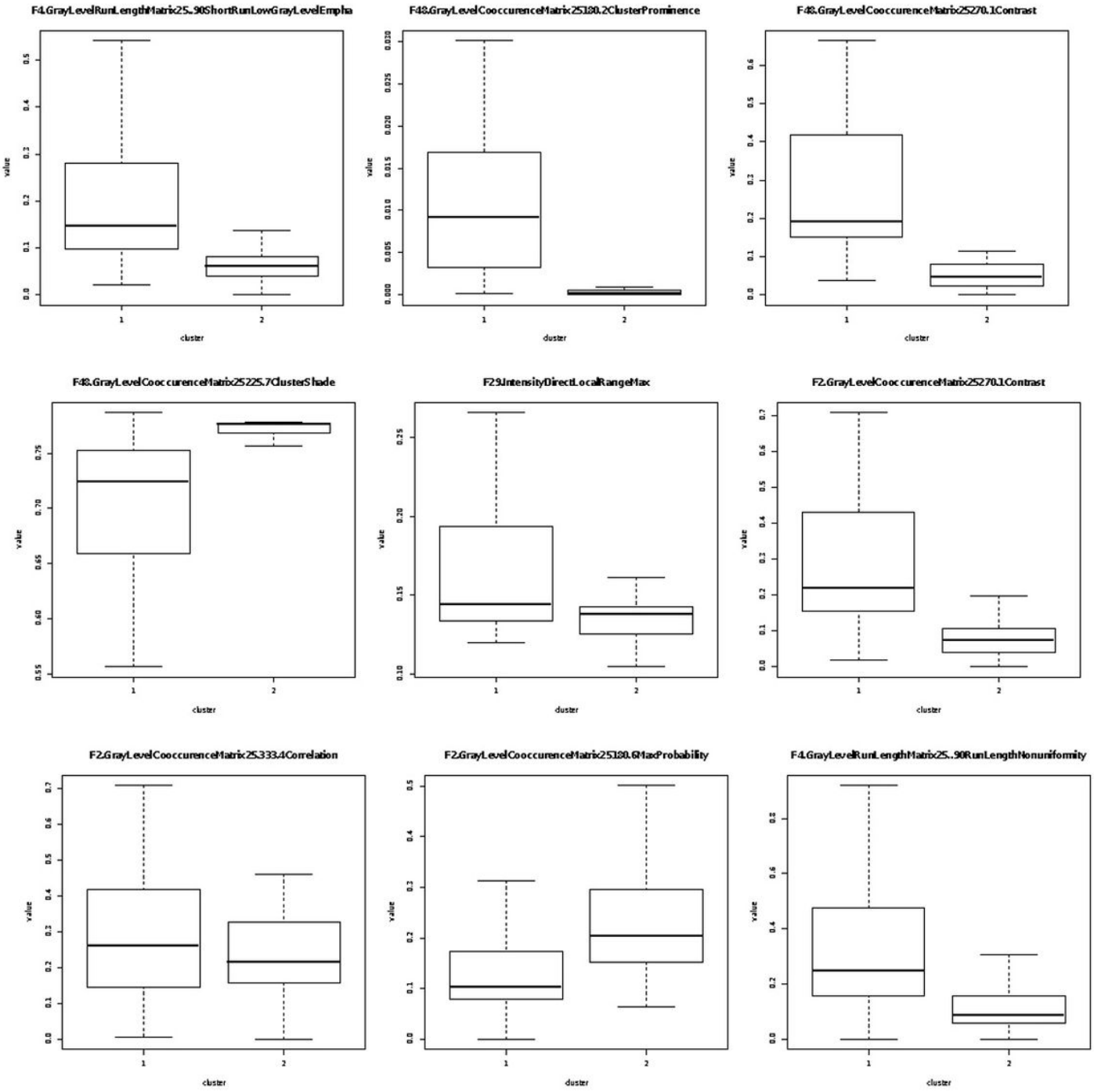


Figure 4

Top Radiomic Features identified by the Random Survival Forest (RSF) for Overall Survival (OS). Boxplots of top 9 features selected using the variable importance from the Random Survival Forest (RSF) over the training data and their distribution within the two clusters identified for Overall Survival (OS). The difference in distribution suggests that these variables can be used in a model to assign cluster labels to test patients.

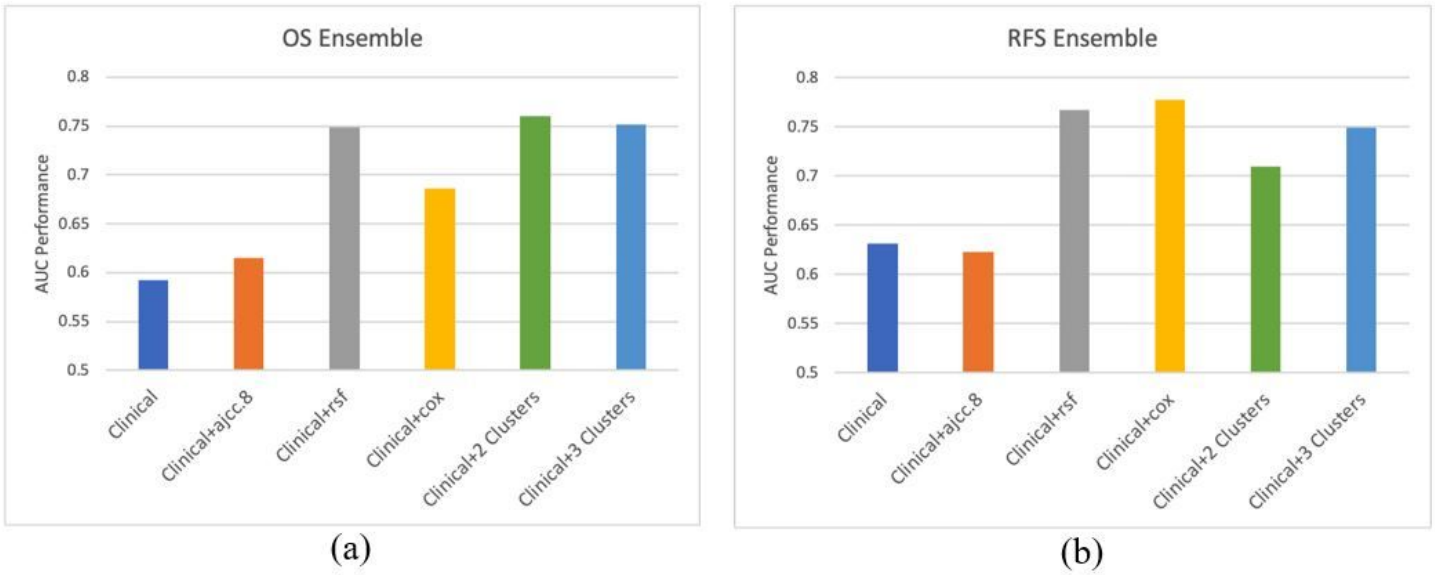


Figure 5

Ensemble model performance over test data. The ensemble model discrimination was evaluated using the AUC metric over the test data for two survival outcomes: (a) Overall Survival (OS) and (b) Recurrence Free Survival (RFS). Comparison is done between a Clinical baseline model using six clinical covariates: age, hpv status, smoking status, T-category, N-category, and therapeutic combination and the models including additional model covariates: AJCC staging (Clinical+ajcc.8), selected radiomic features (Clinical+rsf/+cox), and the proposed cluster labels (Clinical+N Clusters). In all cases, the inclusion of the cluster labels outperforms Clinical and Clinical+ajcc.8 models. For OS, the best performing model is the Clinical+2 Clusters model with AUC:.76. For RFS, the Clinical+3 Clusters reaches AUC:0.75, comparable to the Clinical+rsf/+cox models (AUC:.78) while being a considerably more parsimonious model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- scientificreportsupplementalinformation.docx