IEEE Access
The journal for rapid open access publishing

Received September 20, 2014, accepted October 23, 2014, date of publication November 25, 2014, date of current version December 10, 2014.

Digital Object Identifier 10.1109/ACCESS.2014.2373335

Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features

SAMUEL H. HAWKINS[1], JOHN N. KORECKI[1], YOGANAND BALAGURUNATHAN[2], YUHUA GU[2], VIRENDRA KUMAR[2], SATRAJIT BASU[1], LAWRENCE O. HALL[1], DMITRY B. GOLDGOF[1,2], ROBERT A. GATENBY[2], AND ROBERT J. GILLIES[2]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA
[2]Department of Imaging, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

Corresponding author: S. H. Hawkins (shhawkins@mail.usf.edu)

This work was partially supported by the Radiomics of Non-Small Cell Lung Cancer through the National Institutes of Health under Grant 1U01CA143062-01.

ABSTRACT Nonsmall cell lung cancer is a prevalent disease. It is diagnosed and treated with the help of computed tomography (CT) scans. In this paper, we apply radiomics to select 3-D features from CT images of the lung toward providing prognostic information. Focusing on cases of the adenocarcinoma nonsmall cell lung cancer tumor subtype from a larger data set, we show that classifiers can be built to predict survival time. This is the first known result to make such predictions from CT scans of lung cancer. We compare classifiers and feature selection approaches. The best accuracy when predicting survival was 77.5% using a decision tree in a leave-one-out cross validation and was obtained after selecting five features per fold from 219.

INDEX TERMS Computed tomography, CT 3D texture features, support vector machine, Naive Bayes, decision tree.

I. INTRODUCTION

We explore the idea of radiomics being applied to computed tomography scans of non-small cell lung cancer *(NSCLC)*. Radiomics [1] is the extraction of quantifiable and mineable data from medical images. Here, the focus is on extracting features that can be used to predict whether patient survival time will be long or short.

In this work, computed tomography *(CT)* images of patients from the Moffitt Cancer center were collected. These images were obtained when the patient was diagnosed. The images were used in making the diagnosis. The images all have a different field of view and many have different reconstructions including slice thickness. The variability makes this a challenging data set.

Ganeshan et al. showed that features extracted from CT images of lung tumors correlate with glucose metabolism and stage information [2]. The work by Samala et al. [3] sought the optimum image features to represent lung nodules. Those features were then used in a classification module of a computer-aided diagnosis system. Way et al. [4] tried to distinguish benign nodules from malignant ones using only texture based image features. Lee et al. [5] also performed a detailed study on the usefulness of image features in the classification of pulmonary nodules based on CT-scan images. The work by Zhu et al. [6] showed the effectiveness of a support vector machine in classifying benign and malignant pulmonary nodules. Work has also been done by Al-Kadi and Watson [7] in differentiating between aggressive and non-aggressive malignant lung tumors using texture analysis applied to Contrast Enhanced *(CE)* CT scan images. The use of fractal image features in tumor analysis can be found in the work of Kido et al. [8]. The high level information within CT scans was highlighted by correlating imaging features with global gene expression in hepatocellular carcinoma [9]. They showed that combinations of twenty-eight image features obtained from CT images of liver cancer could reconstruct 78% of the global gene expression profiles.

In this paper we predict survival time from CT images. Section II, contains a description of the data set and the features we used to develop the predictive models. Section III has a discussion of the classifiers. In Section IV, the feature selectors are described in detail, and the results are presented in Section V, followed by conclusions in Section VI.

II. DATA SET AND FEATURE EXTRACTION

In this section, we discuss the data set as well as the methods of image pre-preprocessing, segmentation, and feature extraction. Descriptions of the features are also given. The workflow we used to develop predictive models is represented in Figure 1 and is based on work by Kumar et al. [1].

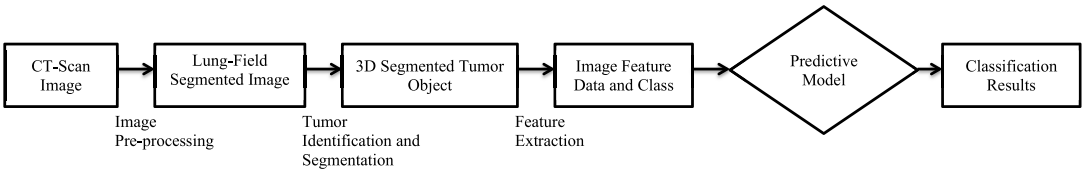


FIGURE 1. Schematic representation of the workflow involved in preparing data for predictive models.

A. DATA SETS

The data set used consisted of de-identified CT-scan images from the Moffitt Cancer Center, Tampa. The images are in the DICOM (Digital Imaging and Communications in Medicine) format. The data set consists of patients with tumor types of Adenocarcinoma and Squamous-cell Carcinoma. This paper focuses on the adenocarcinma patients. CT-scans of 81 adenocarcinoma patients were used for survival time analysis. The slice thickness of the acquired CT-images ranged from 2.5mm to 6mm with an average thickness of 4.75mm. There were 32 cases in stage one, 20 in stage two, 25 in stage three, and 4 cases in stage four. The mean survival time was 879 days. The adenocarcinoma cases were divided into the upper and lower quartiles of survival. The lower quartile consisted of 20 cases surviving from 103 to 498 days with an average survival of 288 days. The upper quartile consisted of 20 cases surviving from 1351 to 2163 days with an average survival of 1569 days. These two classes were chosen in the expectation that their image features would be the easiest to differentiate, provide some information on the possibilities, and the training set is balanced. The class distribution of survival time is as follows:

- patients with a survival time in the highest quartile $[\text{Class1}] = 20$
- patients with a survival time in the lowest quartile $[\text{Class} - 1] = 20$.

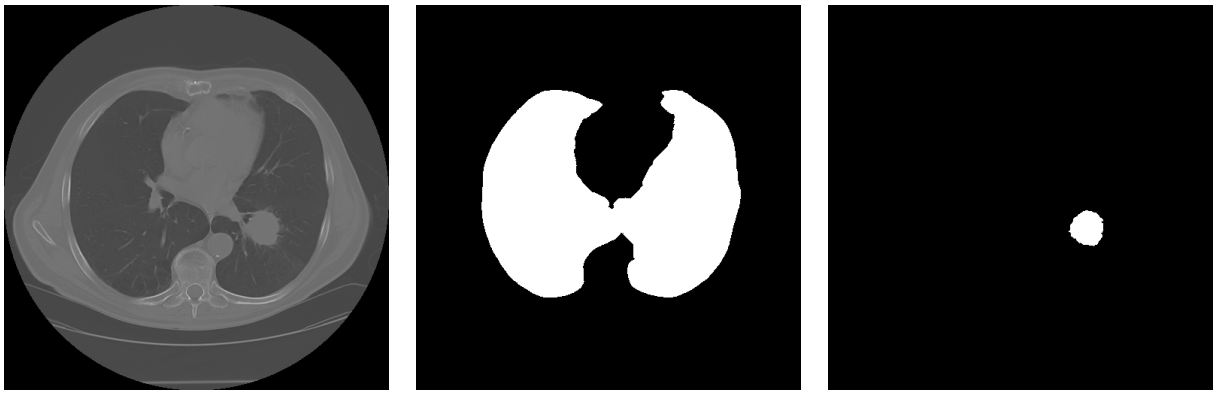
B. IMAGE PRE-PROCESSING

The initial CT segmentation, separating the lung region from the rest of the body, was done using the algorithm provided in the Lung Tumor Analysis (LuTA) software suite of Definiens, [10]. On completion of the lung field segmentation, tumor identification was manually conducted by one of the radiologists at the H. Lee Moffitt Cancer Center or another person with expertise in identifying lung tumors. Upon identification, the tumor was segmented out using the region-growing algorithm developed by Gu et al. [11]. An expert provided the initial seed point for the algorithm. The algorithm finds the tumor boundary across the image sequences. This boundary contains the tumor objects in each slice of the CT-image sequence. Figure 2(a) shows the initial CT image, Figure 2(b) shows the segmentation of the lungs, and Figure 2(c) shows the tumor segmentation after region growing.

C. IMAGE FEATURE EXTRACTION AND FEATURE LIST

In a previous study by Basu et al. [12], a large set of 2D and 3D image features were evaluated for their effectiveness in building a classifier model to distinguish between Adenocarcinoma and Squamous-cell Carcinoma. The study concluded that there was no clear advantage in accuracy between 2D and 3D features, but 3D features simplified constructing classifiers. Thus for this study, only 3D image features were considered.

The image feature extraction algorithms were written in C++ and the executables were embedded into the LuTA software. The image feature extraction was done on only the tumor objects after segmentation by seed growing. The features were normalized from -1 to 1. The major feature types we evaluated are as follows:



(a) (b) (c)

FIGURE 2. Sample CT-image slice. (a) Initial CT image. (b) Lung segmentation. (c) Tumor segmentation.

- Texture features: Co-occurrence Matrices, Laws Features.
- Geometric features: Volume, Rectangular Fit, Compactness, Relative distance measure from pleural wall.
- Intensity based features: mean brightness measure in terms of Hounsfield units (HU), a complete list can be found in the paper by Balagurunathan et al. [13].

III. CLASSIFIERS

The classifiers were selected to test a range of techniques, and to determine those that provide the best predictive accuracy. Below is a brief overview of them.

A. DECISION TREE

The decision tree classifier developed by Quinlan [14] consists of leaves, indicating a class, and branches, specifying a test to be carried out on a single attribute value. Gain-ratio is the performance measure used to decide what test to use at an internal node when building the tree. In a decision tree, a case is classified by starting at the root of the tree and then traversing through the branches until one reaches a leaf. At each branch the case is tested and the outcome decides which subtree to traverse. This process continues until a leaf node is reached and the class is predicted to be the class associated with the leaf. The decision tree used in this study was Weka's J48, [15], which is an implementation of C4.5 release 8 code developed by Quinlan [16]. The decision tree was pruned to make it smaller and more generalizable. The confidence factor for pruning the decision tree was set to 0.25 and the minimum number of cases per leaf was set to 2.

B. RULE BASED CLASSIFICATION

The rule based classifier used was Weka's JRIP, [15], an implementation of the RIPPER algorithm by Cohen [17]. This algorithm consists of two stages. In the grow phase, a rule is extended greedily by adding antecedents until the rule has perfect accuracy. Then in the prune phase the rule is pruned by removing antecedent conditions based on a metric and a pruning data set. Growing and pruning are repeated while there are positive examples or until the error rate exceeds 50%. Finally, rules that would add to the description length are deleted. We used 3 folds, a minimum weight for instances in a rule set to 2.0, 2 optimization runs and a seed of 1 for splitting data into growing and pruning sets.

C. NAIVE BAYES

The Naive Bayes classifier, [18], is designed to be used when features are independent of one another within each class. However, it has been shown that it often works well even when the features are not independent. The Naive Bayes classifier estimates the parameters of a probability distribution given the class, assuming features are conditionally independent. It then computes the posterior probability of a sample belonging to each class and puts the test sample into the class with the largest posterior probability. The assumption of class-conditional independence greatly simplifies the training step. Even though the class-conditional independence between features does not hold for most data sets, this optimistic assumption works well in practice. The classifier labeled Naive Bayes [19] in Weka [15] was used for this work.

D. SUPPORT VECTOR MACHINES

Support vector machines are based on statistical learning theory developed by Cortes and Vapnik [20] and have been shown by Kramer et al. [21], among others, to obtain high accuracy on a diverse range of application domains such as the letter, page, pendigit, satimage, and waveform data sets [22]. SVMs map the input data to a higher dimensional feature space and construct a hyperplane to maximize the margin between classes. A linear decision surface is constructed in this feature space. The hyperplane construction can be reduced to a quadratic optimization problem; subsets of training patterns that lie on the margin were termed support vectors by Cortes and Vapnik [20]. The formulation we used allows for ''errors'' to be on the wrong side of the decision border during training. A cost parameter C is multiplied by the distance the errant example is from the decision boundary. The larger the value of C the larger the penalty applied in the learning process. Parameter tuning of the cost parameter was conducted on training data using a grid search after feature selection. Different kernels, such as a linear kernel, radial basis function kernel, and sigmoid kernel, can be chosen for SVMs. We used the *linear* kernel with a degree of 3. The svm type was set to C-SVC (classification) and the default termination criteria were used. Dehmeshki et al. [23] used support vector machines effectively on CT-scan image data of the lungs in a Computer-Assisted Detection (CAD) system for automated pulmonary nodule detection in thoracic CT-scan images. We used the support vector machine libSVM by Chang and Lin [24].

IV. FEATURE SELECTION

Computed image features can have a high correlation with each other. This property combined with the fact that the number of features available to us was much greater than the number of examples required the investigation of feature selection techniques to improve classification accuracy. Feature selection was done per fold. Leave-one-out cross validation *(LOO)* was conducted on the data. In addition to feature selection, some of the classifiers' models do implicit feature selection. For instance, the decision tree and rule based classifiers subselect features. Also, support vector machines weight features. However, Naive Bayes uses all provided features for classification of the test set. All of classifiers explore all of the features to build models on the training set.

IEEE Access
The journal for rapid open access publishing

TABLE 1. Features chosen by Relief-f feature selection from all of the available features in a leave one out cross validation. Count is how many times the feature was chosen with the maximum being 40, which is once in every fold. Feature name identifies the feature, [13].

Feature Selection	Count	Feature Name
Top 5 Relief-f	40	X8d_3D_Ratio_Free_To_Attached
Top 5 Relief-f	40	X8a_3D_Is_Attached_To_Pleural_Wall
Top 5 Relief-f	40	X8c_3D_Relative_Border_To_PleuralWall
Top 5 Relief-f	40	X8b_3D_Relative_Border_To_Lung
Top 5 Relief-f	40	X3D.Wavelet.decomposition...P1.L2.C14.Layer.1
Top 10 Relief-f	40	X8d_3D_Ratio_Free_To_Attached
Top 10 Relief-f	40	X8a_3D_Is_Attached_To_Pleural_Wall
Top 10 Relief-f	40	X8c_3D_Relative_Border_To_PleuralWall
Top 10 Relief-f	40	X8b_3D_Relative_Border_To_Lung
Top 10 Relief-f	40	X3D.Wavelet.decomposition...P1.L2.C14.Layer.1
Top 10 Relief-f	40	X3D.Wavelet.decomposition...P1.L2.C10.Layer.1
Top 10 Relief-f	39	X3D.Laws.features..E5.S5.R5.Layer.1
Top 10 Relief-f	31	X3D.Laws.features..W5.W5.L5.Layer.1
Top 10 Relief-f	21	X3D.Laws.features..L5.S5.W5.Layer.1
Top 10 Relief-f	5	X3D.Wavelet.decomposition...P1.L2.C9.Layer.1
Top 10 Relief-f	13	avgLRE
Top 10 Relief-f	11	avgSRE
Top 10 Relief-f	7	X3D.Laws.features..L5.L5.S5.Layer.1
Top 10 Relief-f	8	X3D.Laws.features..W5.E5.L5.Layer.1
Top 10 Relief-f	2	X3D.Laws.features..S5.S5.E5.Layer.1
Top 10 Relief-f	2	X3D.Laws.features..W5.S5.L5.Layer.1
Top 10 Relief-f	2	X3D.Wavelet.decomposition...P1.L2.C11.Layer.1
Top 10 Relief-f	4	X3D.Laws.features..S5.S5.W5.Layer.1
Top 10 Relief-f	8	X3D.Laws.features..S5.L5.E5.Layer.1
Top 10 Relief-f	3	X5a_3D_MacSpic_NumberOf
Top 10 Relief-f	2	Histogram.ENERGY.Layer.1
Top 10 Relief-f	1	X3D.Laws.features..R5.E5.L5.Layer.1
Top 10 Relief-f	1	X3D.Laws.features..E5.S5.W5.Layer.1

1) ALL FEATURES

This group includes all 219-image features. No feature selection was performed, thus providing a baseline for the effectiveness of the feature selection techniques.

2) RELIEF-F

The Relief-F algorithm [25]–[27] is a feature evaluator that compares an instance's feature value to the nearest neighbor of both the same and opposite classes. We used a seed of 1, 10 nearest neighbors, and a ranker search. In this work, Relief-F was used to assign ranks to each individual feature. We used the top five and ten features found by the algorithm as shown in Table 1. The top ranked features measure tumor attachment to the wall of the lung.

3) CORRELATION BASED FEATURE SELECTION (CFS)

Correlation based Feature Selection (CFS) searches for features that correlate to a class but do not correlate with each other. The implementation used was found in WEKA, [15] and utilized local prediction. We used a greedy stepwise forward search which generated rankings. CFS discretizes attributes for nominal classes. The features chosen are shown in Table 2. We can see that CFS prefers texture features with a few shape features when compared to the choices of Relief-F. Relief-F focuses on pleural wall attachment type features.

4) TEST-RETEST

Test-retest features were determined by comparing the stability of features generated after two different scans of the same patient fifteen minutes apart [13]. If a feature is repeatable then the two subsequent scans should yield a similar value. The tumor was segmented both manually by a radiologist and with a single click ensemble approach. Different thresholds of correlation were used. Attributes were kept that had a test-retest concordance measured by a concordance correlation coefficient (CCC) of above 0.85, 0.90, and 0.95. At each correlation threshold different attributes were found using the manual and ensemble segmentation methods as well as the intersection of both.

V. RESULTS

This section presents the experimental results of predicting survival.

Table 3 represents the results with the best accuracy and area under the receiver operating curve from a leave-one-out analysis using each classifier. With 40 examples, leave-one-out cross validation is performed by using each

TABLE 2. Features chosen by CFS feature selection from all of the available features in a leave one out cross validation. Count is how many times the feature was chosen with the maximum being 40, which is once in every fold. Feature name identifies the feature, [13].

Feature Selection	Count	Feature Name
Top 5 CFS	40	X3D.Laws.features..W5.S5.R5.Layer.1
Top 5 CFS	40	X3D.Laws.features..W5.S5.W5.Layer.1
Top 5 CFS	13	X3D.Laws.features..R5.S5.S5.Layer.1
Top 5 CFS	13	X3D.Laws.features..S5.S5.W5.Layer.1
Top 5 CFS	14	X3D.Laws.features..W5.S5.S5.Layer.1
Top 5 CFS	28	Longest.Diameter..mm.
Top 5 CFS	26	Short.Axis...Longest.Diameter..mm..
Top 5 CFS	24	Short.Axis..mm.
Top 5 CFS	1	X3D.Laws.features..S5.S5.R5.Layer.1
Top 5 CFS	1	X3D.Wavelet.decomposition...P1.L2.C4.Layer.1
Top 10 CFS	40	X3D.Laws.features..W5.S5.R5.Layer.1
Top 10 CFS	40	X3D.Laws.features..W5.S5.W5.Layer.1
Top 10 CFS	13	X3D.Laws.features..R5.S5.S5.Layer.1
Top 10 CFS	13	X3D.Laws.features..S5.S5.W5.Layer.1
Top 10 CFS	14	X3D.Laws.features..W5.S5.S5.Layer.1
Top 10 CFS	40	Longest.Diameter..mm.
Top 10 CFS	40	Short.Axis...Longest.Diameter..mm..
Top 10 CFS	40	Short.Axis..mm.
Top 10 CFS	40	Mean..HU.
Top 10 CFS	40	StdDev..HU.
Top 10 CFS	28	Volume..cm..
Top 10 CFS	26	X5a_3D_MacSpic_NumberOf
Top 10 CFS	24	X8a_3D_Is_Attached_To_Pleural_Wall
Top 10 CFS	1	X3D.Laws.features..S5.S5.R5.Layer.1
Top 10 CFS	1	X3D.Wavelet.decomposition...P1.L2.C4.Layer.1

TABLE 3. Summary of the highest survival leave-one-out accuracy and AUC results containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. LQ is lower quartile and UQ is upper quartile.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
Decision Tree	**Top 5 Relief-f**	**5**	**77.5%**	**65%**	**90%**	**0.712**
Decision Tree	**Top 10 Relief-f**	**10**	**70%**	**65%**	**75%**	**0.732**
Rules	**All**	**219**	**62.5%**	**65%**	**60%**	**0.729**
Rules	**All Top 5 Relief-f**	**5**	**75%**	**65%**	**85%**	**0.661**
Naive Bayes	**All Top 10 Relief-f**	**10**	**65%**	**55%**	**75%**	**0.52**
Naive Bayes	**Manual & Ensemble test-retest (.85) Top 5 RF**	**5**	**60%**	**45%**	**75%**	**0.64**
SVM	**Manual test-retest (.90) Top 10 Relief-f**	**10**	**65%**	**70%**	**60%**	**0.65**

TABLE 4. Confusion martrix of the top result, 77.5%, using a decision tree classifier with the top five features chosen using Relief-f.

	Predicted Short Survival	Predicted Long Survival
Actual Short Survival	13	7
Actual Long Survival	2	18

subset of 39 examples to do feature selection and build a model using the specified classifier, which is tested on the single held out example. Finally, the accuracy on each held out example is averaged to find the final leave-one-out accuracy.

The highest classification accuracy was 77.5% and was obtained with the decision tree classifier using the top 5 features found by Relief-f. The confusion matrix for this result can be found in Table 4. The highest AUC was with 10 features, chosen by Relief-f at 0.732 for decision trees. For both rule learners and the decision trees there were often few points on the curve. All feature selection was done per fold. CFS had an occasional failure selecting test-retest features and those results are omitted. Also, results that are below 60% accuracy are not listed in the Tables 5-8.

TABLE 5. Survival leave-one-out accuracy results doing further feature selection on test-retest features for decision trees containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. The top accuracy and AUC are in bold.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
Decision Tree	**Top 5 Relief-f**	**5**	**77.5%**	**65%**	**90%**	**0.712**
	Top 10 Relief-f	**10**	**70%**	**65%**	**75%**	**0.732**
	All Top 5 CFS	5	62.5%	95%	30%	0.292
	All Top 10 CFS	10	65%	75%	55%	0.552
	Manual test-retest (.95)	45	60%	95%	25%	0.271
	Manual test-retest (.90) Top 10 Relief-f	10	67.5%	70%	65%	0.562
	Manual test-retest (.90) Top 10 CFS	10	60%	70%	50%	0.435
	Manual test-retest (.85) Top 5 Relief-f	5	62.5%	85%	40%	0.455
	Manual test-retest (.85) Top 5 CFS	5	72.5%	95%	50%	0.488
	Manual test-retest (.85) Top 10 CFS	10	62.5%	70%	55%	0.51
	Ensemble test-retest (.95) Top 5 CFS	5	65%	100%	30%	0.3
	Ensemble test-retest (.95) Top 10 CFS	10	65%	100%	30%	0.3
	Ensemble test-retest (.90) Top 5 CFS	5	62.5%	95%	30%	0.292
	Ensemble test-retest (.90) Top 10 CFS	10	65%	75%	55%	0.524
	Ensemble test-retest (.85) Top 5 CFS	5	62.5%	95%	30%	0.292
	Ensemble test-retest (.85) Top 10 CFS	10	65%	75%	55%	0.524
	Manual & Ensemble test-retest (.90) Top 10 Relief-f	10	65%	70%	60%	0.691
	Manual & Ensemble test-retest (.85) Top 10 Relief-f	10	62.5%	65%	60%	0.68

TABLE 6. Survival leave-one-out accuracy results doing further feature selection on test-retest features for jrip containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. The top accuracy and AUC are in bold.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
Rules	**All**	**219**	**62.5%**	**65%**	**60%**	**0.729**
	All Top 5 Relief-f	**5**	**75%**	**65%**	**85%**	**0.661**
	All Top 10 Relief-f	10	65%	60%	70%	0.598
	Manual test-retest (.90) Top 10 Relief-f	10	62.5%	75%	50%	0.568
	Manual test-retest (.85)	95	62.5%	75%	50%	0.688

TABLE 7. Survival leave-one-out accuracy results doing further feature selection on test-retest features for naive bayes containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. The top accuracy and AUC are in bold.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
Naive Bayes	All Top 5 Relief-f	5	62.5%	45%	80%	0.605
	All Top 10 Relief-f	**10**	**65%**	**55%**	**75%**	**0.52**
	Manual test-retest (.95) Top 5 Relief-f	5	60%	40%	80%	0.458
	Manual test-retest (.90) Top 5 Relief-f	5	60%	45%	75%	0.54
	Manual & Ensemble test-retest (.95) Top 5 Relief-f	5	60%	40%	80%	0.552
	Manual & Ensemble test-retest (.85) Top 5 RF	**5**	**60%**	**45%**	**75%**	**0.64**

TABLE 8. Survival leave-one-out accuracy results doing further feature selection on test-retest features for svm containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. The top accuracy and AUC are in bold.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
SVM	**Manual test-retest (.90) Top 10 Relief-f**	**10**	**65%**	**70%**	**60%**	**0.65**
	Manual & Ensemble test-retest (.90) Top 5 Relief-f	5	62.5%	65%	60%	0.625
	Manual & Ensemble test-retest (.90) Top 10 Relief-f	10	60%	65%	55%	0.6
	Manual & Ensemble test-retest (.85) Top 5 Relief-f	5	62.5%	65%	60%	0.625
	Manual & Ensemble test-retest (.85) Top 10 Relief-f	10	60%	65%	55%	0.6

Tables 5-8 show the results of doing feature selection on the "stable and informative" features from test-retest for the classifiers used here. In these tables we see that while the features selected in the test-retest data sets can be subselected to provide good classifiers, they did not result in the most accurate ones. The data set they come from was more homogenous in scanner type and parameters. Our data set has a different field of view for every

TABLE 9. Survival leave-one-out accuracy results using only volume containing the feature selection method, number of features, average accuracy, lower quartile accuracy, upper quartile accuracy, and the area under the receiver operating curve. The top accuracy and AUC are in bold.

Classifier	Features	#	Avg Accy	LQ Accy	UQ Accy	AUC
Decision Tree	**Volume**	**1**	**45%**	**40%**	**50%**	**0.45**
Rules	Volume	1	32.5%	45%	20%	0.223
Naive Bayes	**Volume**	**1**	**45%**	**60%**	**30%**	**0.388**
SVM	Volume	1	15%	20%	10%	0.15

patient and different slice thicknesses, as well as different scanners.

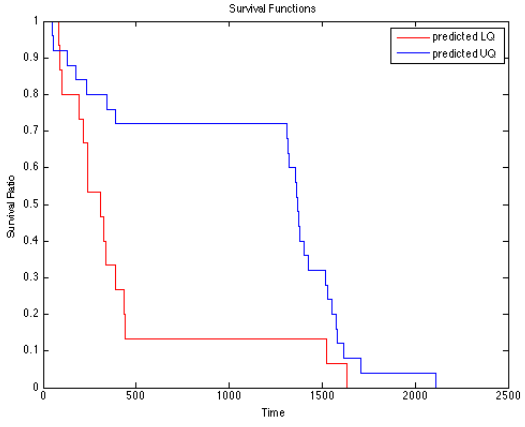


FIGURE 3. Kaplan-Meier curve of the predicted survival classes using our best classifier, a decision tree with five features selected using Relief-f, p = 0.0219.

Figure 3 shows the Kaplan-Meier curve of the survival of the two predicted classes using our best classifier, a decision tree with five features selected using Relief-f. With a p of 0.0219 we reject the null hypothesis that the groups are the same. Thus, the predicted classes are distinct from one another when predicting survival groups.

Table 9 shows the results when training with only volume as a feature. This feature can be useful for differentiating benign from malignant nodules. Here, its accuracy is too low to be useful.

VI. CONCLUSIONS

This is the first study we know of to examine the use of image features from CT scans at the time of diagnosis to predict survival time on a heterogeneous data set. The accuracy of 77.5% is promising and is the highest known accuracy for this problem. This result, using five features chosen with Relief-f, was well above what we were able to achieve using volume alone. The image features from the CT scans may represent phenotypes capable of allowing more accurate predictions than can be made by human analysis alone. The variability of the imaging parameters is a major concern when developing predictive models using, predominantly, image features. If the same field of view and slice thickness were used for all cases, then precision could increase. Clearly, future work requires new stable image features and perhaps an approach using an ensemble of classifiers in which different subsets of features may further improve the accuracy of survival prediction.

ACKNOWLEDGMENT

The authors are grateful to Steven A. Eschrich for his help curating this data.

REFERENCES

[1] V. Kumar *et al.*, "Radiomics: The process and the challenges," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1234–1248, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0730725X12002202

[2] B. Ganeshan, S. Abaleke, R. C. D. Young, C. R. Chatwin, and K. A. Miles, "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: Initial evidence for a relationship with tumour glucose metabolism and stage," *Cancer Imag.*, vol. 10, no. 1, pp. 137–143, Jul. 2010.

[3] R. Samala, W. Moreno, Y. You, and W. Qian, "A novel approach to nodule feature optimization on thin section thoracic CT," *Acad. Radiol.*, vol. 16, no. 4, pp. 418–427, 2009.

[4] T. W. Way *et al.*, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," *Med. Phys.*, vol. 33, no. 7, pp. 2323–2337, 2006. [Online]. Available: http://dx.doi.org/10.1118/1.2207129

[5] M. C. Lee *et al.*, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," *Artif. Intell. Med.*, vol. 50, no. 1, pp. 43–53, 2010.

[6] Y. Zhu, Y. Tan, Y. Hua, M. Wang, G. Zhang, and J. Zhang, "Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography," *J. Digit. Imag.*, vol. 23, no. 1, pp. 51–65, 2010.

[7] O. S. Al-Kadi and D. Watson, "Texture analysis of aggressive and nonaggressive lung tumor CE CT images," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 7, pp. 1822–1830, Jul. 2008.

[8] S. Kido, K. Kuriyama, M. Higashiyama, T. Kasugai, and C. Kuroda, "Fractal analysis of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography: Comparison of bronchioloalveolar cell carcinomas with nonbronchioloalveolar cell carcinomas," *J. Comput. Assist. Tomograph.*, vol. 27, no. 1, pp. 56–61, 2003.

[9] E. Segal *et al.*, "Decoding global gene expression programs in liver cancer by noninvasive imaging," *Nature Biotechnol.*, vol. 25, pp. 675–680, May 2007. [Online]. Available: http://www.stanford.edu/group/OTL/lagan/06343/Kuoetal2007Article.pdf

[10] *Definiens Developer XD 2.0.4 User Guide*, Definiens AG, München, Germany, 2009.

[11] Y. Gu *et al.*, "Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach," *Pattern Recognit.*, vol. 46, no. 3, pp. 692–702, Mar. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312004384

[12] S. Basu *et al.*, "Developing a classifier model for lung tumors in CT-scan images," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2011, pp. 1306–1312.

[13] Y. Balagurunathan *et al.*, "Reproducibility and prognosis of quantitative features extracted from CT images," *Translational Oncol.*, vol. 7, no. 1, pp. 72–87, Feb. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1936523314800107

[14] J. R. Quinlan, "Decision trees and decision-making," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 2, pp. 339–346, Mar./Apr. 1990.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.

[17] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.

[18] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML* (Lecture Notes in Computer Science), vol. 1398, C. Nédellec and C. Rouveirol, Eds. Berlin, Germany: Springer-Verlag, 1998, pp. 4–15. [Online]. Available: http://dx.doi.org/10.1007/BFb0026666

[19] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, 1995, pp. 338–345.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[21] K. A. Kramer, L. O. Hall, D. B. Goldgof, A. Remsen, and T. Luo, "Fast support vector machines for continuous data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 989–1001, 2009.

[22] K. Bache and M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[23] J. Dehmeshki, J. Chen, M. V. Casique, and M. Karakoy, "Classification of lung data by sampling and support vector machine," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (IEMBS)*, Sep. 2004, pp. 3194–3197.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[25] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Workshop Mach. Learn.*, 1992, pp. 249–256.

[26] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learn.*, 1994, pp. 171–182.

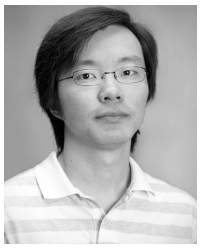
[27] M. Robnik-Sikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 296–304.

SAMUEL H. HAWKINS was born in Tallahassee, FL, USA, in 1982. He received the B.A. and M.S. degrees from Emory University, Atlanta, GA, USA. He is currently pursuing the Ph.D. degree in computer science with the University of South Florida, Tampa, FL, USA.

He has been a Research Assistant with the University of South Florida since 2011.

JOHN N. KORECKI received the B.B.A. in Management Information Systems and the B.S. in Computer Science from the University of Notre Dame. He went on to receive a M.S. in Computer Science from the University of South Florida. His research interests included machine learning and data mining techniques for large feature, small examples data sets with applications to microarray analysis. He is currently an information security software engineer at Verizon developing a large-scale user access data repository.

YOGANAND BALAGURUNATHAN is currently an Applied Research Scientist with the H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. He received the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2001, and the M.E. and B.E. degrees in electronics and communication engineering from Anna University, Chennai, India, and the University of Madras, Chennai, respectively. He was an Assistant Research Engineer with the Genomics Signal Processing Laboratory, Texas A&M University, from 2001 to 2003. He worked on modeling noises effects on microarray technology and applied machine learning methods to genomics in collaboration with the National Human Genome Research Institute, National Institutes of Health, Washington, DC, USA. He was with the Computational Biology/Integrated Cancer Genomics Division, Translational Genomics Research Institute, Phoenix, AZ, USA, as a Senior Research Scientist from 2003 to 2011. He has worked on diverse research fields from imaging, soil science, and life sciences and published his findings in peer-reviewed journals. His research interests include radiology/radiomics, bioinformatics/computational biology, pattern recognition, and nonlinear image/signal processing.



YUHUA GU received the B.S. degree in Applied Mathematics from Fudan University, Shanghai, China in 2001 and the Ph.D degree in Computer science and engineering from University of South Florida, Tampa, Florida, USA in 2009. From 2009 to 2014, he was a Postdoctoral Fellow researcher with H. Lee Moffitt Cancer Center & Research Institute, his research interest includes medical image processing, Radiomics, image segmentation and data mining.

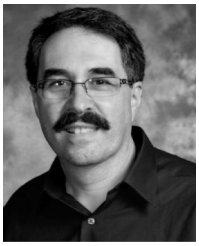
VIRENDRA KUMAR received the M.Sc. degree in biotechnology from IIT Roorkee, Roorkee, India, in 2000, and the Ph.D. degree from the Department of Nuclear Magnetic Resonance, All India Institute of Medical Sciences, New Delhi, India, in 2007.

He was a Post-Doctoral Fellow with the H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA, from 2008 to 2012. Since 2012, he has been an Assistant Professor with the Department of Nuclear Magnetic Resonance, All India Institute of Medical Sciences. He has authored one book chapter, 30 peer-reviewed journals, and over 45 conference articles. His research interests include medical image analysis, radiomics, and biomedical applications of nuclear magnetic resonance and MRI.

Dr. Kumar was a recipient of the ML Wig Gold Medal from the All India Institute of Medical Sciences for the best thesis in 2009 and the Educational Stipend Award from the International Society for Magnetic Resonance in Medicine.

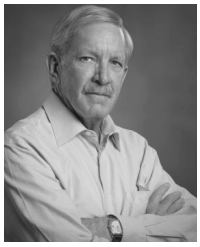
SATRAJIT BASU received the B.Tech. degree in computer science and engineering from the West Bengal University of Technology, Kolkata, India, in 2009, and the M.S. degree in computer science from the University of South Florida, Tampa, FL, USA. He is currently a Data Mining Engineer with SiteWit Corporation, Tampa. His current work at SiteWit Corporation involves using machine learning, and optimization algorithms to maximize the effectiveness of paid search campaigns by automating bid management and keyword discovery. He was a Graduate Research Assistant with the Department of Computer Science, University of South Florida, from 2010 to 2012. His research involved developing classifiers and predictive model frameworks for classification and prognosis of lung tumors using image and clinical features.

LAWRENCE O. HALL (F'03) is a Distinguished University Professor and the Chair of the Department of Computer Science and Engineering at University of South Florida. He received his Ph.D. in Computer Science from the Florida State University in 1986 and a B.S. in Applied Mathematics from the Florida Institute of Technology in 1980. He is a fellow of the IEEE. He is a fellow of the AAAS and IAPR. He received the Norbert Wiener award in 2012 from the IEEE SMC Society. His research interests lie in distributed machine learning, extreme data mining, bioinformatics, pattern recognition and integrating AI into image processing. The exploitation of imprecision with the use of fuzzy logic in pattern recognition, AI and learning is a research theme. He has authored or co-authored over 75 publications in journals, as well as many conference papers and book chapters. He has received over $3M in research funding from agencies such as the National Science Foundation, National Institutes of Health, Department of Energy, and NASA.



MOFFITT DMITRY B. GOLDGOF (F'07) is an educator and scientist working in the area of biomedical image analysis, video processing, pattern recognition and bioengineering. He is currently a Professor and Associate Chair, Department of Computer Science and Engineering at the University of South Florida. Dr. Goldgof has graduated 22 Ph.D. and 43 MS students, published over 80 journal and 200 conference papers, 20 books chapters and edited 4 books (citations impact: h-index 42, g-index 75). Professor Goldgof is a Fellow of IEEE and a Fellow of IAPR and is currently serving on the IEEE Press Editorial Board. Dr. Goldgof is an Associate Editor for IEEE Transactions on Systems, Man and Cybernetics and for International Journal of Pattern Recognition and Artificial Intelligence.

ROBERT A. GATENBY, MD is the Chair of the Department of Radiology at H. Lee Moffitt Cancer Center and Co-Director of the Cancer Biology and Evolution Program. He joined Moffitt in 2008 from the University of Arizona where he was Professor, Department Radiology and Professor, Department of Applied Mathematics since 2000. He received the B.S.E. in Bioengineering and Mechanical Sciences from Princeton University and the M.D. from the University of Pennsylvania in 1977. He completed his residency in radiology at the University of Pennsylvania where he served as chief resident. Bob remains an active clinical radiologist specializing in body imaging. While working at the Fox Chase Cancer Center after residency, Bob perceived that cancer biology and oncology were awash in data but lacked coherent frameworks of understanding to organize this information and integrate new results. Since 1990, most of Bob's research has focused on exploring mathematical methods to generate theoretical models for cancer biology and oncology. His current modeling interests include: the tumor microenvironment and its role in tumor biology, evolutionary dynamics in carcinogenesis, tumor progression and therapy, and information flow in living systems and its role in maintaining thermodynamic stability.



ROBERT J. GILLIES is currently Chairman of the Department of Cancer Imaging and Metabolism; Director of the Center of Excellence in Cancer Imaging and Technology; Vice-chair for Research in the Department of Radiology; and Scientific Director of the Small Animal Imaging Lab (SAIL) at the H. Lee Moffitt Cancer Center and Research Institute in Tampa, FL.

In addition to authoring over 200 peer-reviewed manuscripts, Dr. Gillies has received numerous local, national, and international awards for his teaching and research, including; Researcher of the Year-2012 (Moffitt Cancer Center), the Furrow Award for Innovative Teaching (U. Arizona), the Yuhas Award for Radiation Oncology Research (U. Penn), the TEFAF professorship (U. Maastricht), and the award for Distinguished Basic Scientist of 2009 from the Academy of Molecular Imaging.

Dr. Gillies' vision for the Moffitt imaging initiative includes development of new applications to diagnose, predict and monitor therapy response using noninvasive imaging. This work spans from molecular and chemical, from animal studies to human clinical trials and patient care. Dr. Gillies also leads a post-doctoral/resident training program in cancer imaging. His research is focused on functional and molecular imaging of cancer, specifically with an emphasis on the use of imaging to inform evolutionary models of carcinogenesis and response to therapy.