FULL PAPER

Non-invasive classification of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features

1USMAN BASHIR, FRCP, MD, 2BHAVIN KAWA, FRCP, 1MUHAMMAD SIDDIQUE, PhD, 3SZE MUN MAK, FRCP, 3ARJUN NAIR, FRCP, MD, 4EMMA MCLEAN, FRCPath, 5ANDREA BILLE, FRCS, PhD, 6VICKY GOH, FRCP, MD and 6GARY COOK, FRCP, MD

1Cancer Imaging Department, School of Biomedical Engineering and Imaging Sciences, King’s College London, London, UK
2Department of Radiology, Maidstone Hospital, Hermitage Lane, Maidstone, UK
3Department of Radiology, Guy’s Hospital, Great Maze Pond, London, UK
4Department of Pathology, Guy’s Hospital and St Thomas’ NHS Foundation Trust, Westminster Bridge Rd, Lambeth, London, UK
5Department of Thoracic Surgery, Guy’s Hospital, Great Maze Pond, London, UK
6PET Imaging Centre and the Division of Imaging Sciences and Biomedical Engineering, King’s College London, UK

Address correspondence to: Dr Usman Bashir
E-mail: drusmanbashir@gmail.com

Objective: Non-invasive distinction between squamous cell carcinoma and adenocarcinoma subtypes of non-small-cell lung cancer (NSCLC) may be beneficial to patients unfit for invasive diagnostic procedures or when tissue is insufficient for diagnosis. The purpose of our study was to compare the performance of random forest algorithms utilizing CT radiomics and/or semantic features in classifying NSCLC.

Methods: Two thoracic radiologists scored 11 semantic features on CT scans of 106 patients with NSCLC. A set of 115 radiomics features was extracted from the CT scans. Random forest models were developed from semantic (RM-sem), radiomics (RM-rad), and all features combined (RM-all). External validation of models was performed using an independent test data set (n = 100) of CT scans. Model performance was measured with out-of-bag error and area under curve (AUC), and compared using receiver-operating characteristics curve analysis on the test data set.

Results: The median (interquartile-range) error rates of the models were: RF-sem 24.5 % (22.6 – 37.5 %), RF-rad 35.8 % (34.9 – 38.7 %), and RM-all 37.7 % (37.7 – 37.7). On training data, both RF-rad and RF-all gave perfect discrimination (AUC = 1), which was significantly higher than that achieved by RF-sem (AUC = 0.78; p < 0.0001). On test data, however, RM-sem model (AUC = 0.82) out-performed RM-rad and RM-all (AUC = 0.5 and AUC = 0.56; p < 0.0001), neither of which was significantly different from random guess ( p = 0.9 and 0.6 respectively).

Conclusion: Non-invasive classification of NSCLC can be done accurately using random forest classification models based on well-known CT-derived descriptive features. However, radiomics-based classification models performed poorly in this scenario when tested on independent data and should be used with caution, due to their possible lack of generalizability to new data.

Advances in knowledge: Our study describes novel CT-derived random forest models based on radiologist-interpretation of CT scans (semantic features) that can assist NSCLC classification when histopathology is equivocal or when histopathological sampling is not possible. It also shows that random forest models based on semantic features may be more useful than those built from computational radiomic features.

INTRODUCTION

Non-small cell lung cancers (NSCLCs) comprise 85% of all primary lung malignancies.1 Of these, approximately 60% are adenocarcinomas (ADCA) and 35–40% are squamous cell carcinomas (SCCA), with large cell cancers accounting for less than 5%.1 Conventionally, ADCA and SCCA are differentiated by histopathological examination of haematoxylin & eosin-stained slides. ADCAs, depending upon the predominant pathologic subtype, may exhibit lepidic, glandular, papillary or micropapillary, or solid sheet-like architecture. SCCAs are characterized by the presence of keratinization, pearl formation, and intercellular bridges.2 Frequently, NSCLC is diagnosed on sputum cytology or clinical and radiological features, but adequate
tissue is not available to perform histological subtyping and molecular analysis, requiring a multidisciplinary approach for decision-making.² Although curative options for both NSCLC subtypes are similar—either surgical or with stereotactic body radiotherapy (SABR—the two subtypes differ in prognosis and choice of targeted agents.³ Hence, an accurate non-invasive test for NSCLC classification could serve as a valuable alternative for prognostication and choosing targeted agents in patients unsuitable for surgical resection.

Radiomics and machine learning (ML) are becoming increasingly popular in imaging research.⁴ Radiomics involves computational analysis of a greyscale image to derive features (e.g. mean, mode, kurtosis, and skewness) which are expected to quantify the tumour pathophysiology.⁵ ML is the task of using radiomics and other relevant variables (e.g. age, sex, and air bronchogram) in suitable computational algorithms (e.g. random forests or logistic regression) to infer clinically relevant information, e.g. tumour subtype. CT radiomics has been shown to be moderately to highly accurate in predicting NSCLC subtype, with reported performance of 68–90%.⁶–⁸ However, despite the potential of radiomics in changing imaging paradigms,⁹ widespread acceptance of radiomics is hindered by largely unmet challenges surrounding variable reproducibility, procedure standardization, and biologic explanation of used variables.⁴⁻⁹,¹⁰

Semantic features, i.e. features derived from subjective interpretation of CT images by a radiologist, have been shown to be related to tumour subtype and histopathology in numerous independent studies.¹¹⁻¹⁷ Air-bronchogram and ground-glass opacification are more common in ADCA, whereas cavitation and spiculation are more common in SCCA.¹⁶⁻¹⁷ To our knowledge however, despite these well-known associations, semantic features have not been modelled in ML algorithms to predict tumour subtype and therefore help clinical decision making in a quantitative manner. Furthermore, no studies have compared or combined radiomic features with semantic features (e.g. air bronchogram and cavitation) in differentiating ADCA from SCCA.

We hypothesized that multivariate predictive models combining the strengths of semantic and radiomic features could yield potentially higher accuracy in NSCLC classification than either class of variables alone. Such non-invasive classification would benefit patients for whom an adequate histopathological subtyping cannot be obtained. Therefore, the objective of this study was to develop and compare NSCLC classification models based on semantic features, radiomic features, and a combination of both.

**METHODS AND PATIENTS**

**Patient population**

The training data set comprised patients referred to a single institution as follows: we identified pre-treatment CT scans of pathologically proved NSCLC patients referred to our tertiary care centre from January 1, 2011 to December 31, 2015. Patients were excluded if it was not possible to accurately determine tumour boundaries on CT, e.g. due to adjacent atelectasis. The final data set comprised 106 studies (42 SCCA, 64 ADCA; Figure 1). The independent validation cohort (n = 100) comprised 65 ADCAs and 35 SCCAs downloaded from the Cancer Imaging archive, subsampled with respect to ADCAs to ensure balanced proportions.¹⁸⁻²⁰ Local ethics committee waived informed written consent for this retrospective study of anonymised data.

**Imaging**

Imaging of patients in the training data set was performed on one of three Philips scanners: MX8000, Brilliance iCT 256, or Brilliance 40 (Philips Medical Systems, Best, Netherlands). Patients were imaged in the supine position at full inspiration. Scanning parameters were as follows: detector collimation: 0.625–0.75 mm; rotation time: 0.5–0.75 s; tube voltage: 120 kVp; tube current: 34–229 mAs. 100–150 ml iopromide 300 (300 mg I/mL Ultravist, Bayer Pharma, Berlin, Germany) was administered intravenously at a rate of 2–4 ml s⁻¹ after a 30–70 s delay.

**Semantic features**

Two thoracic radiologists (AN and MM, with 14 and 9 years’ experience, respectively), blinded to histopathological diagnosis, independently recorded nine nodule semantic features (Table 1) and two background parenchymal features, i.e. emphysema (present or absent) and airway thickening (present or absent).¹¹,¹²,¹⁻²⁶

---

**Figure 1.** Patient inclusion workflow in our study for training and validation data sets. ADCA, adenocarcinoma; NOS, not otherwise specified; NSCLC, non-small-cell lung cancer; SCCA, squamous cell carcinoma; TCIA, The cancer imaging Archive.
Discrepant findings were resolved by consensus. Annotation of the validation data set was performed by a separate blinded reader, UB (10 years’ radiology experience), using the same descriptions.

Radiomic features
Tumours were delineated by UB open-source software ITK-Snap (v. 3.4.0; Supplementary Material 1).27 From the segmented volumes-of-interest, 756 radiomic features were derived using an in-house feature extraction tool developed in MATLAB (Release 2016b, The MathWorks, Inc., Natick, MA). Highly correlated redundant features (showing pairwise correlation coefficient >0.8; n = 641) were removed to yield a final set of independent 115 radiomic features.

Random forest model development and validation
In this study, we used random forests for machine learning. Random forests are known for their high performance and generalizability.28 Here we present a summary of random forest model development; technical details are provided in the supplemental data.

A random forest model is a group of a large number of decision trees, e.g. 2000. The name “random” alludes to the fact that each split of an individual decision tree is developed from a random subset of input variables. Each member tree is also trained on a slightly different variation of the data-set by using bootstrap sampling, i.e. sampling with replacement, whereby several cases are sampled more than once and others omitted altogether (labelled “out-of-bag” [OOB] samples). Since the OOB samples have not been used in training the particular tree, they are used for internal validation, and the proportion of misclassified cases in the OOB sample serves as a performance metric: OOB error.

After training of all 2000 decision trees is complete, a new case may be classified as ADCA if >50% trees classify it as ADCA, or SCCA otherwise.

We developed three random forest classifiers using the training data set: One classifier comprising semantic variables only (RF-sem), one comprising radiomic features only (RF-rad), and one comprising both semantic and radiomic features (RF-all). Model validation was performed on the independent validation cohort.

<table>
<thead>
<tr>
<th>Clinical feature</th>
<th>ADCA</th>
<th>SCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years, mean (range, SD)</td>
<td>69 (40.2–84.75, 10.2)</td>
<td>70.8 (52.35–85.54, 8.1)</td>
</tr>
<tr>
<td>Sex (M : F)</td>
<td>32 : 32</td>
<td>24 : 18</td>
</tr>
<tr>
<td>Smokers</td>
<td>65.6% (n = 42)</td>
<td>71.4%(n = 30)</td>
</tr>
<tr>
<td>T1a</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>T1b</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>T2a</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>T2b</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>T3</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>T4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>N0</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>N1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N2</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M0</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>M1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

ADCA, adenocarcinoma; SCCA, squamous cell carcinoma; SD = standard deviation.
Statistical analysis

R v. 3.3.2 was used for statistical analysis. Continuous variables were reported as means and standard deviations. For descriptive analysis, differences between ADCA and SCCA were determined using Wilcoxon ranked sum test for continuous variables and using Fisher’s exact test for categorical variables. Interobserver agreement between the two radiologists with regards to semantic variables was measured with Cohen’s $\kappa$ test and summarised as estimated weighted $\kappa$ scores along with their 95% CIs. A $p$-value cut-off of 0.05 was used to determine statistical significance.

The performance of random forest models was reported in terms of two metrics: The OOB error of random forest models was reported as the error rate of decision trees during internal validation. The second metric—area under curve (AUC)—served as the performance metric of fully trained models and was reported separately for training and validation data. We used two metrics instead of one to illustrate both the robustness of individual trees (OOB error) and that of the forest as a whole (AUC). Both metrics are related, and an ideal classifier should have both a low OOB error and a high AUC.

Since our random forests used large numbers of variables, we also measured the importance of individual variables in the training data set using the “mean decrease in accuracy” (MDA) metric, i.e. decrease in classifier accuracy by removing the variable in question. The higher the MDA of a variable the more important the variable is. A variable with MDA of zero has no association with the outcome (tumour subtype) and there is no decrease in classifier accuracy if that variable is removed. Variables with low but non-zero MDA are still useful since random forests by design work well when individual variables are weakly related to the outcome, and mitigate their weak association by pooling them into a robust final classifier.

RESULTS

The mean interval between pathologic diagnosis and CT chest imaging was 21 days (range 5–41 days). Patients were aged from 40.3 to 85.5 years (median: 71.4 years), with similar gender proportions (50 females: 56 males). There were no significant differences between patients with ADCA vs SCCA in terms of age ($p = 0.6$), smoking ($p = 0.67$), or gender (0.55) (Table 2).
Of the 13 tested semantic variables, three were significantly more common in ADCAs, i.e., air bronchogram ($p < 0.0001$), ground-glass component ($p = 0.0006$), and satellite nodules ($p = 0.004$). Cavitation was present in only 9 of the 106 total cases, of which 8 were SCCAs ($p = 0.002$). Table 3 describes the frequencies of semantic variables in both NSCLC subtypes.

Comparison of random forest models

The semantic random forest (RF-sem) performed equally well on training and test data sets with AUC of 0.78 and 0.82 respectively (Figure 2). The radiomics-only and combined models gave consistent good performance on training (B; AUC 0.78) as well as test data (B; AUC 0.82). AUC, area under the curve; RF, radiofrequency.

In terms variable importance, air bronchogram (MDA = 0.039), ground-glass component (MDA = 0.023), and cavitation (MDA = 0.019) were the top-ranking semantic variables, whereas tumour location, spiculation, and tumour margins did not have any discriminatory value. Of the radiomic variables, the highest ranking variables were grey-level size-zone matrix (GLSZM) short zone low intensity emphasis (GLSZM-SZLIE; MDA = 0.005), co-efficient of variation (MDA = 0.004), and neighbourhood grey-tone difference matrix (NGTDM) coarseness (MDA = 0.003). Variable importance of semantic features and top 10 ranking radiomic features (total = 756) is given in Table 4.

**DISCUSSION**

We developed three NSCLC classification models. RF-sem utilised semantic features obtained by consensus between two thoracic radiologists from training data and by a separate radiologist, from the validation data. RF-rad was based on computer-aided extraction of radiomic features from CT images of NSCLCs, whereas RF-all was a combination of semantic and radiomic features. RF-sem performed well on both training and validation data despite both data sets having been annotated by separate radiologists, indicating the robustness of random forest models developed with semantic features to inter-observer variability. RF-rad and RF-all gave perfect predictions on training
features such as spiculation and central location. In our cohort, are known to be more frequent in SCCA, in addition to semantic variables including older age, male gender, and smoking history.

We found several semantic features highly predictive of NSCLC subtype (Table 5), of which air-bronchogram, ground-glass component, cavitation, and satellite nodules ranked highest in terms of discriminatory capability (Table 4). Our findings regarding the relative proportions of the various semantic features support previously reported trends, with a few differences. Several clinical variables including older age, male gender, and smoking history are known to be more frequent in SCCA, in addition to semantic features such as spiculation and central location. In our cohort, none of these variables were significantly different between ADCAs and SCCAs and did not make a substantial contribution to the classifier.

The most important radiomic features in our study were GLSZM-SZLIE (MDA = 0.005), coefficient of variation (MDA = 0.004), and NGTDM coarseness (MDA = 0.003). The biologic counterparts of these features are poorly understood; here we attempt an intuitive explanation of what these features might represent in tumour CT images: The GLSZM, described originally for texture characterisation of cell nuclei, quantifies image heterogeneity in terms of zones of contiguous voxels sharing the same grey level intensity. A relatively homogeneous tumour would have large zones of voxels sharing similar grey level intensity and vice versa. The derived quantity GLSZM-SZLIE, as the name implies, would be expected to be high in tumours with heterogeneous distribution of low grey-level (e.g. ground-glass density) voxels. NGTDM coarseness, originally tested on various natural (e.g. pebbles, grass) and synthetic materials (e.g. cloth), would be high in tumours exhibiting similar intensities in neighbouring voxels with a low spatial rate of change in voxel intensities. In other words, they would comprise clusters of similar-intensity voxels which would stand out against the background and give a 'coarse' appearing texture to the tumour. Coefficient of variation (ratio of standard deviation over mean) is a first-order statistical texture feature which is high in tumours exhibiting high variation in grey-level intensities and low mean intensities. All three features were slightly more common in ADCAs vs SCCAs in our cohort.

A few authors have previously explored radiomics in NSCLC classification: In their proof of concept study, Basu et al trained a classifier (accuracy: 68%) on CT-derived radiomic features from 74 cases of NSCLC. Their study focused on differentiating the efficacy of 2D radiomic features versus 3D radiomic features and presented a comparison of various model categories including random forests, support vector machines, decision trees, and nearest neighbours. Their best model accuracy of 68% was obtained by employing all 215 features in a leave-one-out cross-validation scheme. However, the authors did not report the best performing variables and a comparison of various model categories including random forests, support vector machines, decision trees, and nearest neighbours. Two recent studies done by Wu et al. and Zhu et al. reported higher performance of radiomics-models (AUC 0.72 and 0.9 respectively). Other than that neither study compared radiomic features with semantic features, the most important difference between our study and either two is that the subset of highest performing radiomic features is different in all three studies. It is possible that since there are hundreds of radiomic features with majority inter-correlated, some of the different high-ranking features might merely be variations of the same feature. A second possibility is that some of the radiomic models developed by other authors may have overfit, as seen in our study, although Wu et al used an external validation cohort making this unlikely in their study. Overfitting is a common design problem in ML studies, especially in studies with a large number of variables with respect to cases and lack of external validation cohort. Radiomics is doubly challenged in gaining widespread acceptance due to the common use of hundreds of variables and issues surrounding reproducibility, although efforts are underway to standardise radiomics.

Our study has several potential limitations: Because this was a CT study, we could not completely eliminate the possibility of including small regions of normal tissue, e.g. opacification due...
to adjacent atelectasis. However, we minimised such cases by excluding lesions that were difficult to delineate from adjacent collapsed lung. As a result, there may have been an under representation of centrally located SCCAs because such tumours were frequently inseparable from adjacent atelectasis. Central location is a known feature of SCCAs and including more centrally located tumours, expected to be majority SCCA, may have improved model performance.33 Secondly, as in most radiomics studies, our original radiomic feature space comprised a large number ($n = 756$) of features derived from CT scans with varying data acquisition parameters, especially those obtained from T1CA. Radiomic features are variable in terms of reproducibility and are dependent on tumour segmentation and image post-processing steps.27 Hence, we believe that future studies using a more refined selection of radiomic features, especially features engineered specifically for chosen classification tasks, may provide more useful results.

CONCLUSIONS

Our study showed that non-invasive classification of NSCLCs using semantic features is possible and can be done with good accuracy (AUC: 0.82) using machine learning algorithms. However, CT-scan radiomic features performed poorly on independent validation data (AUC 0.5 and 0.56 for RF-tex and RF-all respectively), despite perfect classification on test data, and may be unsuitable for this task.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from the King's College London /University College London Comprehensive Cancer Imaging Centres funded by Cancer Research UK and Engineering and Physical Sciences Research Council in association with the Medical Research Council and the Department of Health(C1519/A16463); from the Wellcome/EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z).

REFERENCES

5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016; 278: 563–77. doi: https://doi.org/10.1148/radiol.2015151169


