Computers in Biology and Medicine 136 (2021) 104752

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass CT radiomics signature

Zahra Khodabakhshi [a], Shayan Mostafaei [b,c], Hossein Arabi [d], Mehrdad Oveisi [e,f], Isaac Shiri [d], Habib Zaidi [d,g,h,i,*]

[a] Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Science, Tehran, Iran
[b] Department of Biostatistics, School of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran
[c] Epidemiology and Biostatistics Unit, Rheumatology Research Center, Tehran University of Medical Sciences, Tehran, Iran
[d] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva 4, Switzerland
[e] Department of Computer Science, University of British Columbia, Vancouver BC, Canada
[f] Comprehensive Cancer Centre, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom
[g] Geneva University Neurocenter, Geneva University, Geneva, Switzerland
[h] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[i] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
NSCLC
Histopathology
Radiomics
CT
High-dimensional multinomial classification

ABSTRACT

Objective: The aim of this study was to identify the most important features and assess their discriminative power in the classification of the subtypes of NSCLC.
Methods: This study involved 354 pathologically proven NSCLC patients including 134 squamous cell carcinoma (SCC), 110 large cell carcinoma (LCC), 62 not other specified (NOS), and 48 adenocarcinoma (ADC). In total, 1433 radiomics features were extracted from 3D volumes of interest drawn on the malignant lesion identified on CT images. Wrapper algorithm and multivariate adaptive regression splines were implemented to identify the most relevant/discriminative features. A multivariable multinomial logistic regression was employed with 1000 bootstrapping samples based on the selected features to classify four main subtypes of NSCLC.
Results: The results revealed that the texture features, specifically gray level size zone matrix features (GLSZM), were the significant indicators of NSCLC subtypes. The optimized classifier achieved an average precision, recall, F1-score, and accuracy of 0.710, 0.703, 0.706, and 0.865, respectively, based on the selected features by the wrapper algorithm.
Conclusions: Our CT radiomics approach demonstrated impressive potential for the classification of the four main histological subtypes of NSCLC, It is anticipated that CT radiomics could be useful in treatment planning and precision medicine.

1. Introduction

Lung cancer is the most frequently diagnosed and first cause of death from cancer in the world [1]. According to global cancer statistics, it is the most common cancer type among men and women with incidence, mortality, and five-year survival rate of 11.6%,18.4%, and 15%, respectively [1–3]. Lung cancer is clinically categorized into two main groups, including small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC is more prevalent, accounting for almost 85% of all lung cancer cases [4]. It has three main histological subtypes: Adenocarcinoma (ADC, 40%), squamous cell carcinoma (SCC, 25%–30%), and large cell carcinoma (LCC, 10%–15%). Another subtype is "not otherwise specified", denoted as NOS, which has none of the specific characteristics of the aforementioned subtypes [5,6].

In precision medicine, therapeutic decisions including surgery, chemotherapy, immunotherapy, or radiotherapy, are taken based on the cancer stage, histologic and genetic characteristics of the patient's tumor [6]. For example, the recommended treatments for patients with advanced epidermal growth factor receptor (EGFR) mutated adenocarcinoma include erlotinib, gefitinib, and afatinib [7]. Conversely, for squamous cell lung cancer cases, therapeutic options include platinum doublet chemotherapy and immune-checkpoint inhibitors (ICIs) [6,7].

* Corresponding author. Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).

https://doi.org/10.1016/j.compbiomed.2021.104752
Received 20 May 2021; Received in revised form 21 July 2021; Accepted 5 August 2021
Available online 8 August 2021

Moreover, prognosis and reoccurrence rates are significantly different between ADC and SCC groups wherein the reoccurrence risk is dramatically higher in non-SCC groups [8]. In this light, the determination of the histopathology of NSCLC before treatment is of crucial importance.

Currently, pathological analysis by means of biopsy or preparation of tissue sections after surgery is the first-line reference to the identification of NSCLC subtypes [9]. However, these methods have some inherent limitations and challenges [9]. First of all, they involve invasive procedures and the preparation of the pathological tissue section is costly and time-consuming. Moreover, the biopsy is a snap-shot of a small part of the whole tumor and given the heterogeneous nature of lung tumors, especially ADC type, this will not reflect the entire characteristics of the tumor. Moreover, these methods are not recommended for patients at advanced inoperable stages [9].

In routine clinical practice, Computed tomography (CT) is commonly used for lung cancer diagnosis due to its high sensitivity for the detection of small nodules [10]. Subtypes of NSCLC have different phenotypic characteristics on CT images. For instance, cavitation and speculation are more associated with SCC, whereas ground-glass opacification and air-bronchogram are significant CT imaging signatures of ADC [11,12]. In this regard, sometimes these phenotypic signs may become hidden or mixed/confused with the underlying structures in CT images. Hence, the visual discrimination of NSCLC subtypes is a real challenge for radiologists as it is highly dependent on their knowledge, experience, and analytical skills. In this light, an alternative accurate and noninvasive approach for the histological classification of NSCLC would be crucial in clinical practice.

Recent studies have demonstrated the promising potential of radiomics in the histological classification of cancerous lesions [13–16]. Radiomics refers to the extraction of high throughput features from medical images that bear information related to tumor pathobiology and the creation of a minable high dimensional database [17–19]. A radiomics pipeline was employed to develop descriptive and diagnostic models to associate/relate image phenotypes to gene expression [17, 20–23]. Previous studies focused mainly on the differentiation between ADC and SCC based on the extracted features from CT images [14,15,24] and all of them confirmed the capability of radiomics analysis in NSCLC subtypes classification. Wu et al. [14] implemented several classifiers for the classification of ADC and SCC using CT radiomics features. Junior et al. [24] extracted 2277 radiomics features from 68 CT images with proven malignant lung lesions and investigated the performance of classifiers on oversampled, undersampled, and original unbalanced data with radiomics features, clinical features, and a combination of both. Zue et al. [15] investigated the performance of five selected radiomics features in the classification of NSCLC histological subtypes. Han et al. [25] used PET/CT images of 867 patients presenting with ADC and 552 patients with SCC and investigated the optimal classification model among radiomics-based, machine learning, and deep learning algorithms. The authors reported that the best performing machine learning model (support vector machine) achieved area under the receiver operating characteristics (AUROC) curve of 0.863 and accuracy of 0.792.

Overall, all of these studies demonstrated the potential of radiomics in the classification of NSCLC subtypes. These studies considered only two subtypes of NSCLC, wherein the multi-subtype classification of NSCLC is of crucial importance since patients with other subtypes, including LCC and NOS require different treatment strategies. To the best of our knowledge, only two studies performed multi-subtypes classification of NSCLC. Patil et al. [13] investigated the performance of support vector machines with two groups of features, including non-texture features (volume and first-order) and texture features. According to their results, texture features had higher potential for multi-subtype classification of NSCLC in comparison with non-texture features. Lui et al. [26] developed a hybrid technique to classify the four main subtypes of NSCLC. The proposed model, referred to as SLS, combined three algorithms, namely synthetic minority oversampling method, L2,1-norm minimization and, support vector machine into a single model.

In this study we aimed to identify the most significant features and evaluate their discriminative power in the classification of the subtypes of NSCLC patients. To deal with the minimal-optimal problem, which refers to identifying minimal set of the most relevant and informative features, we used multivariate adaptive regression splines (MARS) a powerful algorithm which is also appropriate for high-dimensional data modeling [27] and wrapper algorithm as a novel feature selection algorithm for finding all relevant variables [28].

2. Materials and methods

The different steps of the methodology followed in the current study are summarized in Fig. 1.

2.1. Data preparation and image segmentation

The clinical studies used in this work were obtained from the publicly available NSCLC radiomics-genomics dataset on the cancer imaging archive (TCIA) [29]. The dataset included 422 patients with NSCLC. During the data curing procedure, the dataset was reduced to 354 patients owing to the absence of histopathological information and/or poor image quality. The final dataset included 134 SCC, 110 LC, 62 NOS, and 48 ADC cases. The 3D volumes of primary gross tumor volumes (GTV) were manually delineated by a radiation oncologist.

2.2. Image preprocessing and feature extraction

In this work, 1433 radiomics features were extracted from each tumor volume. The features were extracted using the PyRadiomics package, an open-source Python package for the extraction of radiomics features from medical images in which the feature definition is compliant with the image biomarker standardization initiative (IBSI) [30,31]. The quantitative features were divided into three subgroups, namely (i) morphological features, (ii) first-order statistical features, and (iii) texture features. The value of radiomic features depends highly on the number of gray levels and voxel size [32,33]. Therefore, we used voxel size resampling and gray level discretization to harmonize voxel sizes, reduce noise and variability of radiomic features. Images were resampled into $1 \times 1 \times 1$ m^3 isotropic voxel size and discretized into 64 bins prior to feature extraction.

Prior to feature extraction, the original images were decomposed into 8 wavelet decomposition (HHH, HHL, HLH, LHH, LLL, LHL, HLL, LLH). Moreover, Laplacian of Gaussian filters (LOG) with different sigma values (0.5–5 with step 0.5) were applied to the original images. Since morphological/shape feature extraction does not require any preprocessing, we extracted these features directly from the original images. Texture features play a pivotal role in medical image analysis. These features are calculated by considering the relationship between the intensity at each voxel and those at the neighboring voxels in various directions [34]. Since these features represent intensity distribution within the tumor region, they could be affected by various imaging parameters. Hence, it is crucial to apply mathematical transformations to the image before extracting these features since transformation and filtering methods, such as the wavelet transformation and LOG filters can reduce noise and enhance edge detection, thus improving the representation of specific structures on images. Several studies have indicated that the calculated texture features based on wavelet decomposed images have significant diagnostic or prognostic values [14,35]. Therefore, we applied different filters and transformations on the images prior to feature extraction. Sixty texture features and 19 first-order statistical features were calculated for each wavelet decomposition and LOG filtered images. Fig. 2 illustrates schematically the feature extraction process. Details of the extracted features are provided in

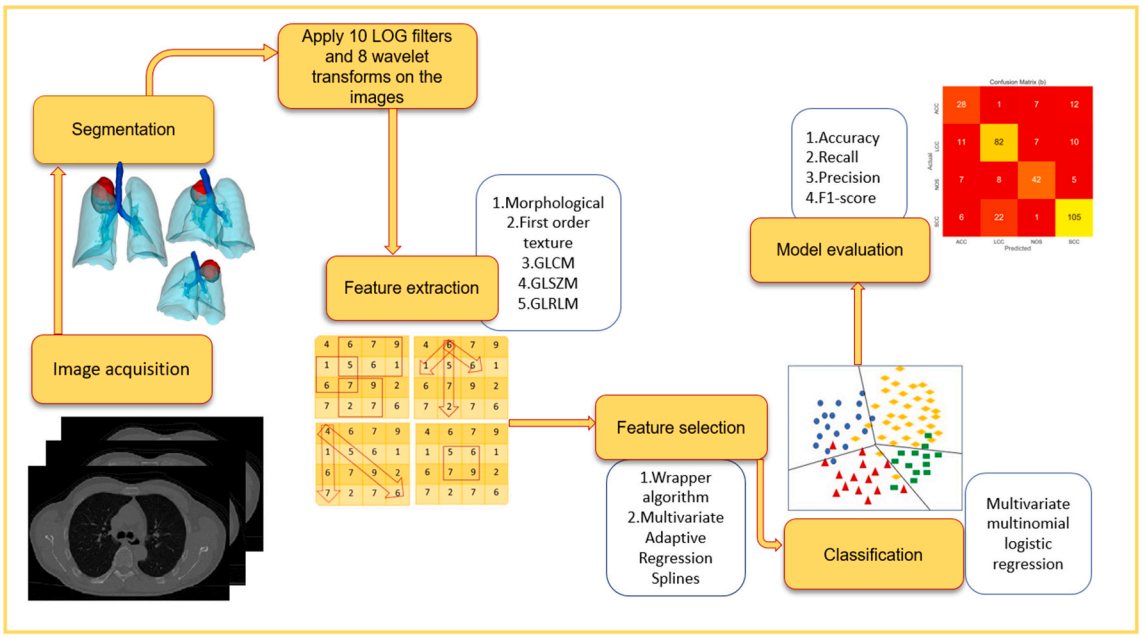

Fig. 1. The Radiomics pipeline employed in this study. GLCM: Gray Level Co-Occurrence Matrix; GLRLM: Gray Level Run Length Matrix; GLSZM: Gray Level Size Zone Matrix.

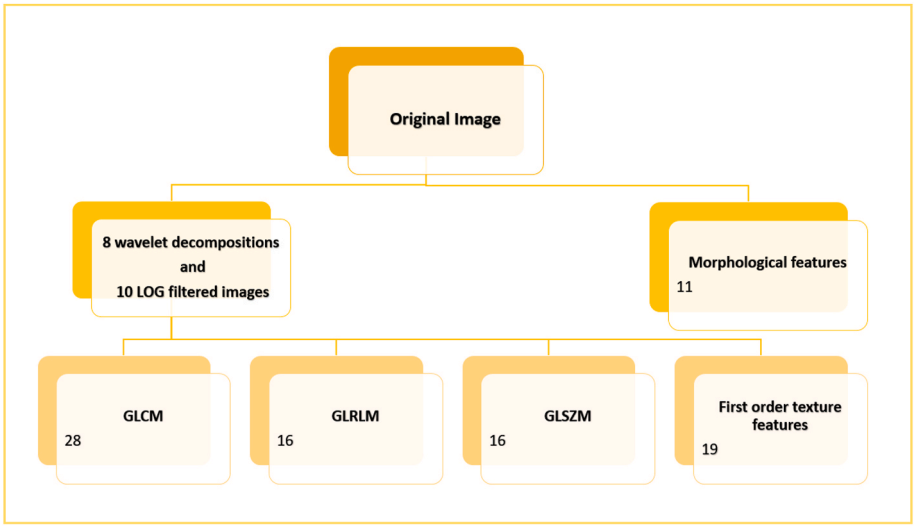

Fig. 2. Schematic illustration for the feature extraction process. Eleven morphological features were extracted from the original image. Sixty texture and 19 first-order statistical features were extracted from each wavelet decomposition and LOG filtered images.

supplementary Table 1.

2.3. Feature selection

Two novel classifiers were employed to remove irrelevant and redundant features which do not contribute to the accuracy of a prediction model, namely the wrapper random forest feature selection algorithm for finding all relevant features by "Boruta" R package and multivariate adaptive regression splines method by "earth" R package. In this ensemble (bagging method) wrapper algorithm, feature importance index was calculated for all features, then the significant features were selected consistently [36]. In addition, MARS, a piece-wise linear and non-linear segments regression modeling of high-dimensional data [27], has exhibited superior performance over the traditional classification and regression algorithms in finding optimal variable transformations and interactions in high-dimensional complex data structure [37]. This algorithm employs a bootstrap procedure that estimates the performance of classification. This algorithm assess the importance value of each feature/variable/predictor by removing the variable and evaluating the decrease in calculated generalized cross-validation criteria. Regarding the previous examination on both real and simulated studies, MARS not only has higher classification accuracy compared to support vector machine, random forest, and elastic-net methods [38] but also cause lower Type II errors and its associated risks [37].

2.4. Multinomial classification

After feature refinement, the selection of the most important CT features using a multinomial logistic regression model as a dimension reduction algorithm, was used to assess the predictive probability of the selected features. To this end, the stepwise variable selection procedure

was employed using a 2-tailed alpha with a value of 5% for insertion and deletion criteria. In addition, 1000 bootstrapping samples were taken to reduce the variance of the classification power indices. The multinomial logistic regression model via bootstrapping was applied by "nnet" R package [39]. P-values were adjusted by Benjamini and Hochberg method and a p-value < 0.05 was considered as statistically significant threshold. In this study, the whole dataset was fed into feature selection algorithms to select the most relevant features. Due to lack of external validation set and small sample size, classification was performed and validated on selected features using 5-fold cross-validation with bootstrap resampling that was repeated 1000 times.

3. Results

3.1. Feature selection

From 1433 extracted features, 14 features were consistently selected based on their overall importance value. Fig. 3 presents the top 9 selected features by the wrapper algorithm and the top 11 selected features by multivariate adaptive regression splines. According to Figs. 3, 6 out of 14 consistently selected features were common between the two algorithms where 3D GLSZM-gray level non-uniformity normalized of log sigma2, wavelet-HLH-GLSZM-small feature area emphasis, and wavelet-LHL-GLSZM-zone entropy had the highest importance value (40%, 38%, and 36%, respectively). Regarding Fig. 3, most of the selected features belong to the GLSZM category whereas only two features fall under the first-order category.

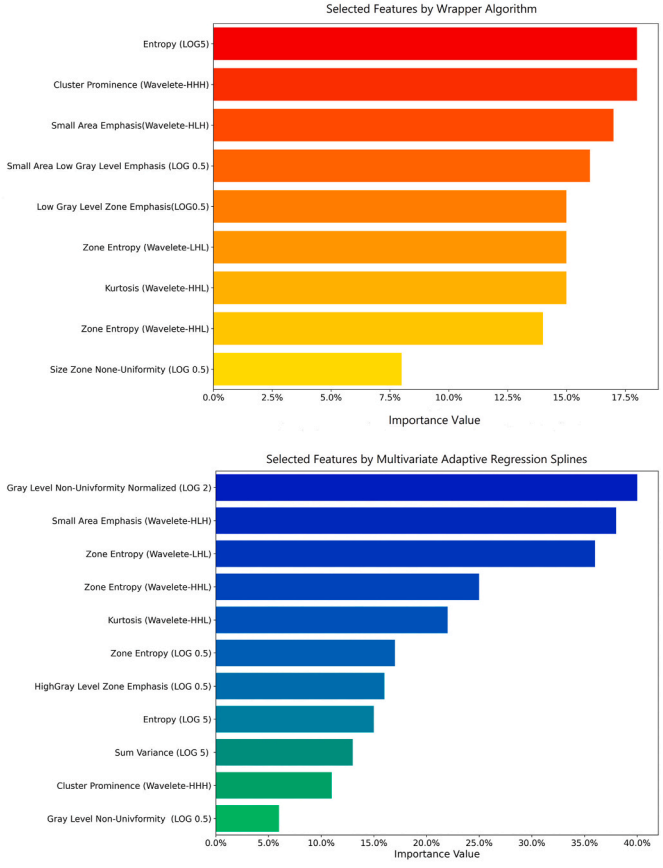

Fig. 3. Bar plot of feature importance for the selected features using the Wrapper algorithm and multivariate adaptive regression splines.

3.2. Classification

Both groups of selected features were used for subtypes classification by multivariable multinational logistic regression with 1000 bootstrapping samples. In Fig. 4, the confusion matrix shown in (Fig. 4a) represents the classification results based on the wrapper algorithm feature selection whereas the confusion matrix (Fig. 4b) represents results based on multivariate adaptive regression splines. We used 4 classification metrics to evaluate the classifier's performance, which could be calculated based on the confusion matrix. These metrics include precision, recall, F1-score, and accuracy. For each class, precision is the true predicted positive divided by total predicted positive, whereas the sensitivity (or recall) is the ratio of true predicted positive of a class to the actual positive cases of the same class. Accuracy equals the ratio of correctly predicted cases to the total predictions, whereas F1-score is defined as the weighted average of precision and sensitivity (recall).

The classification performance of our models is shown in Table 1, wherein for the features selected by the wrapper algorithm, the model performance reached the average accuracy, F1-score, precision, recall, and AUC of 0.865, 0.706, 0.710, 0.703, and 0.747 respectively. The same metrics for the second model are 0.862, 0.697, 0.699, 0.679, and 0.723 respectively. According to these results, the performance of the classifiers for both sets of features is approximately similar. The classifier worked poorly on ADC classification and was mainly due to the limited number of ADC cases in comparison with the other 3 subtypes (Table 1). The ROC curves of different classes are also provided in Fig. 5 (the ADC class was considered as the reference class).

Based on the coefficients and adjusted P-values of multinomial logistic regression models in Table 2, log sigma5, Kurtosis from first-order in wavelet (HHL), wavelet-HHL_glszm_ZoneEntropy and wavelet-HLH_glszm_SmallAreaEmphasis have the most significant effects on the prediction of the subtypes of NSCLC patients in the wrapper algorithm with the model's goodness of fit index of pseudo $R^2 = 0.159$. Moreover, log sigma2 and the mentioned features have the most significant effects on the prediction of subtypes of NSCLC patients in the MARS model with a model's goodness of fit index of pseudo $R^2 = 0.173$.

4. Discussion

Even though using the underlying and fundamental medical image features, e.g. tumor volume or diameter are common for cancer staging [40], medical images contain a lot more valuable information beyond what could be seen visually. Recent advances in pattern recognition algorithms and availability of the accurately categorized datasets have facilitated the process of medical image task-oriented feature identification. In this context, radiomics utilizes automated algorithms to transform medical images into high-dimensional minable sets of quantitative features and analyzes the phenotypic characteristics to assist clinical decisions.

In this work, we extracted a large number of radiomics features from three sets of CT images, including original images, wavelet decompositions and LOG filtered images, followed by applying two different advanced methods for an effective feature selection. MARS, as an adaptive nonparametric regression, is a piecewise linear and nonlinear segments flexible regression modeling for high-dimensional data (e.g. radiomics). MARS is capable of fitting complex and nonlinear relationships between the dependent variable and predictors and performing traditional classification and regression algorithms, such as logistic regression, neural networks, and support vector machine to find/detect the optimal variable transformations and interactions in high-dimensional and complex data structures [37,41,42]. In addition, Boruta or Random Forest with the wrapper method, as a bagging ensemble machine learning method, has emerged as an efficient and robust algorithm that can handle feature selection problems in high-dimensional data [36]. One of the main advantages of this wrapper

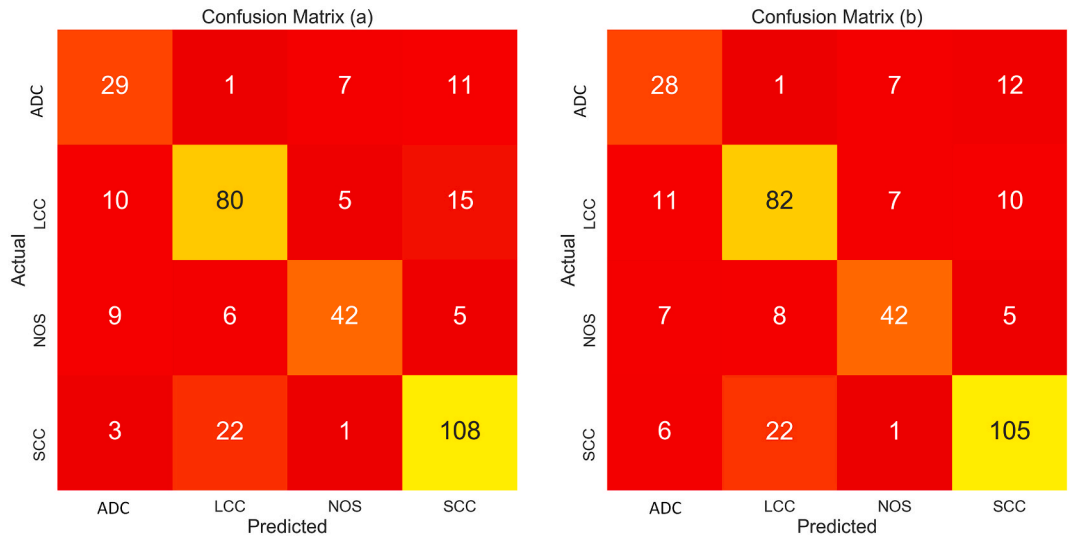

Fig. 4. Results of classification for the different subtypes of NSCLC showing: (a) Confusion matrix obtained from the wrapper algorithm and (b) Confusion matrix obtained from the multivariate adaptive regression splines.

Table 1
The classification power indices (SD) by multivariable multinomial logistic regression with 1000 bootstrapping samples based on the methods of feature selection.

Method	Subtype	Precision	Recall/ Sensitivity	F1-score	Accuracy	Specificity	AUC (95%CI)
Wrapper algorithm	**Adenocarcinoma**	0.568 (0.022)	0.604 (0.016)	0.585 (0.018)	0.884 (0.066)	0.928 (0.061)	Ref.
	Large cell	0.733 (0.051)	0.727 (0.050)	0.730 (0.076)	0.833 (0.077)	0.881 (0.053)	0.67 (0.58–0.76)
	NOS	0.763(0.068)	0.677 (0.032)	0.717 (0.048)	0.906 (0.109)	0.955 (0.077)	0.69 (0.59–0.79)
	Squamous cell carcinoma	0.776 (0.10)	0.805 (0.111)	0.791 (0.125)	0.838 (0.131)	0.859 (0.071)	0.88 (0.81–0.96)
	Average/total	0.710	0.703	0.706	0.865	0.906	0.747
Multivariate Adaptive Regression Splines	**Adenocarcinoma**	0.538 (0.032)	0.583 (0.033)	0.560 (0.06)	0.875 (0.072)	0.921 (0.062)	Ref.
	Large cell	0.725 (0.049)	0.745 (0.065)	0.735 (0.053)	0.833 (0.066)	0.873 (0.055)	0.65 (0.56–0.74)
	NOS	0.736 (0.082)	0.677 (0.052)	0.705 (0.063)	0.901 (0.123)	0.948 (0.075)	0.65 (0.55–0.75)
	Squamous cell carcinoma	0.795 (0.093)	0.783 (0.112)	0.789 (0.131)	0.841 (0.129)	0.877 (0.076)	0.87 (0.79–0.94)
	Average/total	0.699	0.697	0.697	0.862	0.905	0.723

algorithm is its computation time which is shorter in comparison with other algorithms such as regularized Random Forest and recursive feature elimination algorithms for high dimensional dataset such as gene selection in the genomics [43]. Based on previous studies using large and multicenter dataset, Boruta was the most accurate model compared with Lasso, Elastic Net regularized regression, gradient boosted feature selection, and regularized Random Forest [44]. Lastly, a multivariable multinominal logistic regression model was implemented to classify four main subtypes of NSCLC. The results showed that our models were able to predict NSCLC histological subtypes and proved the potential of radiomics to quantify and extract phenotypic characteristics and predict the histological nature of the lesions.

In this work, we used multinomial logistic regression as a classifier. However, a number of studies implemented modern approaches, including those using deep learning techniques [25,45]. The motivation behind not using this approach was the small sample size. For complex problems, e.g multiclass classification, a larger dataset is required to train and define the weights of a deep learning architecture. Previous deep learning-based radiomic studies focusing on the classification of NSCLC subtypes [25,45], included more than 800 patients, which is significantly higher than the size of our dataset. It should also be stated that using simple models with fewer parameters has the advantage of generalizability keeping in mind that the best model is the simplest model with fewer assumptions [46].

The number and quality of selected features play a pivotal role in radiomics analysis and could have a huge impact on the overall performance of a classifier. One way to increase the number of features is via applying different filters on the images. In this work, we used the wavelet decomposition and LOG filtered images, wherein a total number of 1433 features were extracted, which is significantly higher in comparison with the total extracted features in previous studies [15,26,47,48]. Moreover, recent studies have emphasized the benefits of features extracted from filtered images, specifically LOG filtered images, for the task of tumor subtypes differentiation [49–51]. In our study, 12 out of 14 selected features belonged to the category of texture features which correspond to the heterogeneity of tumor volume, thus reflecting the importance of these types of imaging biomarkers. The feature with the highest importance value (40%) was Gray Level Non-Uniformity Normalized from 3D GLSZM in LOG sigma 2 mm. This feature is the normalized distribution of zone counts over gray values. Its lower values indicate equally distributed zone counts [30]. Due to the high importance value of this feature, as demonstrated in this study, it could be

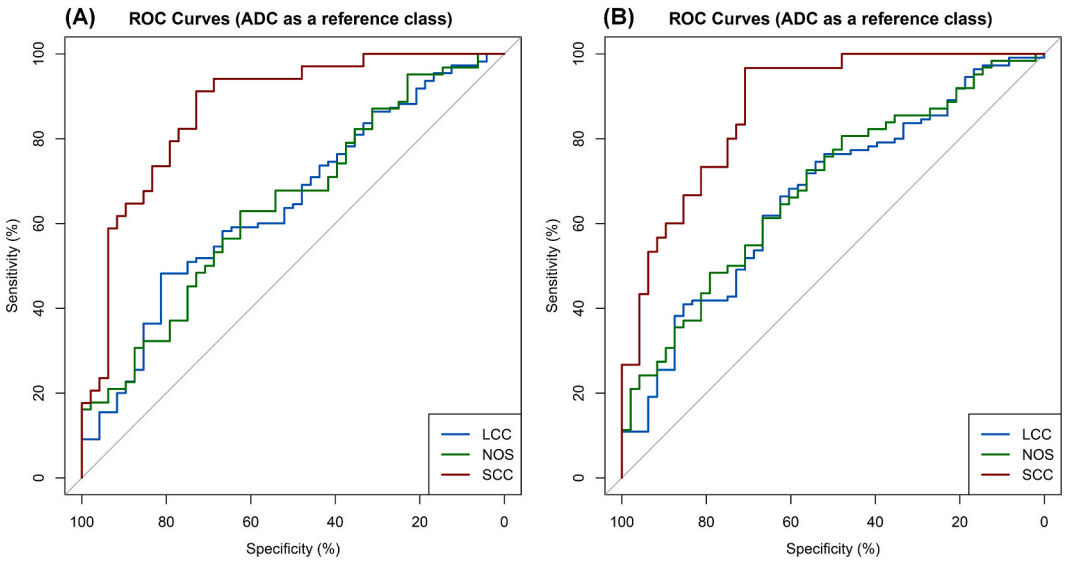

Fig. 5. ROC curves of the classifier performance for the selected features by (A) MARS algorithm and (B) Wrapper algorithm.

Table 2
Multivariable multinomial logistic regression for the selected features by "mnlogit" R package and the model's goodness of fit index, P-value of Wald Chi-square test, Adj. P-value: P-value adjusted by Benjamini and Hochberg method. SE= Standard of Error.

Method	Selected Variables			Adj. P-Value	Coefficient (SE)	Pseudo R^2
	Image preprocessing	**Parent feature**	**Feature**			
Wrapper Algorithm	Log-sigma 0.5 mm	3D-GLSZM	Small Area Low Gray Level Emphasis	0.009	−8.05 (3.09)	0.159
	Wavelet-HHL	First order	Kurtosis	0.001	0.29 (0.018)	
	Wavelet-HHL	GLSZM	Zone Entropy	0.007	−0.95 (0.351)	
	Wavelet-HLH	GLSZM	Small Area Emphasis	0.001	2.36 (0.187)	
	Wavelet-HHH	GLCM	Cluster Prominence	0.199	−3.96 (3.087)	
	Wavelet-LHL	GLSZM	Zone Entropy	0.049	−0.495 (0.30)	
	Log-sigma 0.5 mm	3D-GLSZM	Low Gray Level Zone Emphasis	0.021	0.042 (0.018)	
	Log-sigma 0.5 mm	3D-GLSZM	Size Zone Non-Uniformity	0.068	0.003 (0.002)	
	Log-sigma 5 mm	3D First order	Entropy	0.039	−0.79 (0.45)	
Multivariate Adaptive Regression Splines	Log-sigma 0.5 mm	3D-GLSZM	Zone Entropy	0.002	1.79 (0.571)	0.173
	Wavelet-HHL	First order	Kurtosis	0.001	0.27 (0.022)	
	Wavelet-HHL	GLSZM	Zone Entropy	0.008	−0.94 (0.353)	
	Wavelet-HLH	GLSZM	Small Area Emphasis	0.001	2.05 (0.188)	
	Wavelet-HHH	GLCM	Cluster Prominence	0.328	−2.66 (2.72)	
	Wavelet-LHL	GLSZM	Zone Entropy	0.774	−0.10 (0.36)	
	Log-sigma 5 mm	3D First order	Entropy	0.046	−1.66 (0.98)	
	Log-sigma 2 mm	3D-GLSZM	Gray Level Non-Uniformity Normalized	0.003	1.99 (0.682)	
	Log-sigma 0.5 mm	3D GLSZM	High Gray Level Zone Emphasis	0.006	−15.58 (5.64)	
	Log-sigma 5 mm	3D GLCM	Sum Variance 2	0.027	1.77 (0.921)	
	Log-sigma 0.5 mm	3D GLSZM	Gray Level Non-Uniformity	0.132	0.003 (0.002)	

used as an indicator for NSCLC subtypes classification.

Previous studies mainly focused on the classification of two main subtypes of NSCLC i.e. ADC and SCC. Wu et al. [14] investigated 24 feature selection methods and 3 classification methods on their dataset containing 198 pathologically proven NSCLC cases in the training cohort and 152 cases in the test cohort. In their study, the Naive Baye's classifier achieved the highest performance in terms of AUC (0.72) with 5 selected features. Three out of five selected features belong to first-order statistical features, while two other features belong to GLRLM and GLCM. In a study by Zhu et al. [15] a radiomic signature consisted of five radiomic features was constructed in order to differentiate between ADC and SCC. Radiomic features were selected using LASSO logistic regression model. According to their results, the radiomic signature achieved an AUC of 0.893, a sensitivity of 0.828, and a specificity of 0.9 which is a fairly good result.

These studies demonstrated the strong potential of radiomics in the differentiation between ADC and SCC [14,15,25,47,52]. However, in real life problems, these models are not practical since two other subtypes of NSCLC have been overlooked in those studies and their models would fail to properly differentiate between the different histopathological subtypes, which is crucial for the optimization of treatment strategies. In this light, this study has the merits of comprehensiveness and generalizability considering all subtypes of NSCLC which is a more complex and challenging problem. Only two studies reported on the classification of four subtypes of NSCLC using the SMOTE technique to balance their dataset [13,26]. The first study reported the results of ten-fold cross-validation and did not test their model on a validation dataset [13]. Their model reached an accuracy of 0.88 on the validation set using ten-fold cross-validation. In addition, their validation set of 10-fold cross-validation contained data generated using the SMOTE method. Therefore, the performance of their model should be examined using an external dataset for an objective comparison. In the second study, 20% of the dataset was set aside for model validation [26]. They trained their model with and without oversampled data and reported the results of classification on the test set for both trained models. The average accuracy, F1-score, recall, and precision of

their model on the test set, without using SMOTE for the train data were 0.67, 0.67, 0.67, and 0.68, respectively. Synthetic Minority Oversampling Technique (SMOTE) is a data augmentation method in which the minority class is oversampled by synthesizing new examples from the existing examples [53]. However, our model without using the oversampling technique reached a higher average value for those metrics (0.865, 0.706, 0.703, and 0.71, respectively). Using the SMOTE method, their model achieved good results for both train and test set. However, for the test set, they selected a balanced dataset and the number of each class was almost equal: 17 ADC and NOS cases, and 18 LCC and SCC cases. Nevertheless, in real life clinical scenario, the data are not balanced and it is crucial to test the model on an imbalanced dataset as well. Moreover, oversampling techniques, like SMOTE, have some drawbacks, with overfitting being the most common one [54]. Due to the lack of an external validation set and the small sample size, specifically for ADC class, we used bootstrap resampling for model evaluation. During this procedure, the model was trained on a sample set and evaluated on a sample not included in the training set. The whole procedure was repeated 1000 times. Hence, the results are robust and comparable to those reported in previous studies. In our study, the model based on selected features by the Wrapper algorithm achieved an average accuracy, F1 score, recall, and precision of 0.865, 0.706, 0.703, and 0.71 respectively.

The major obstacle in our study was that our training dataset was imbalanced with a limited number of samples for certain types of NSCLC. As such, the discriminative features associated with certain subtypes of NSCLC (such as ADC) may be overlooked/overshadowed by the dominant features from other subtypes with a larger number of samples. A limited number of samples in a dataset introduces bias and variance in the performance of a classifier [55]. Since most conventional classification algorithms are designed for datasets with a balanced distribution, imbalanced data distribution imposes challenges for both feature selection and model training. In these problems, the classifiers are biased toward major classes (with a larger number of subjects) and classes with a smaller number of samples (in our study, the ADC class) are likely to be misclassified or regarded as noise or outlier [56]. Various approaches were devised to address the problem of training with an imbalanced dataset, including conventional resampling methods, such as SMOTE or cost-sensitive classifiers. However, having a balanced data distribution does not necessarily lead to optimal model performance and other factors, such as training sample size and class complexity (class overlapping) also affect the performance of a classifier [57].

Decoding the underlying and discriminative feature from a non-uniform dataset is highly challenging, particularly for classes with smaller sample sizes. According to a study conducted by Japkowicz et al. [58], increasing the training sample size would decrease the error rate of imbalanced class classification. Another problem considered in our study is class overlapping or class complexity. Separating the minority class from the majority classes is one of the major obstacles, which complicates model training as well as feature selection. When class overlapping is present in feature space, features would not properly represent a specific class and may exhibit low levels of specificity [56]. The aforementioned issues were the root cause of sub-optimal performance of the proposed model for the classification of the ADC subtype. However, a number of approaches including features selection methods were devised to address specifically these issues for imbalanced datasets. The Wrapper method is one of the powerful feature selection methods for these sort of problems [59]. This approach explores the feature subsets space using a learning algorithm to estimate the classification accuracy based on which each feature could be included or eliminated from the feature subsets [56]. This idea was implemented in this work through resampling/bootstrap processing and the results reported in Table3 demonstrated that the selected features improved the overall performance of the classifier.

Our study suffer from some other limitations. First of all, we only used radiomics features of CT images. However, other imaging modalities, such as PET and MRI may boost the performance of machine learning models for histological subtype classifications [25]. Although MRI is not routinely used in the clinic for lung cancer diagnosis and characterization owing to challenges associated with visual assessment, a recent study confirmed the potential of multimodal MRI-based radiomics for discrimination of NSCLC subtypes [60]. The model achieved an AUC of 0.82 for the validation set, which is comparable to CT or PET-based radiomic model for the classification of NSCLC subtypes. Besides, some studies demonstrated that using semantic features e.g nodule location, cavitation speculations, and clinical features, such as age, smoking history, and body mass index can help the differentiation between NSCLC subtypes [48,61]. Another shortcoming of our study, like most radiomics studies, is that our study was retrospective, which bears inherent challenges. Factors, such as image reconstruction techniques, scanning parameters, and tumor delineation methods can strongly affect radiomics features [62–67]. These factors cannot be controlled in retrospective studies and may decrease model performance.

5. Conclusion

Our model accuracy for multi-class and unbalanced classification of NSCLC was reasonable compared to previous studies reported in the literature. However, other classification metrics e.g. precision and recall were not good enough for ADC subtype evaluation. This is one of the limitations of our study owing to insufficient data for this class. A larger dataset could potentially increase the classification accuracy of the proposed models for this subtype. Moreover, the approaches developed for binary classification of the imbalance dataset are not optimal for multi-class classification. Further research and development efforts should be carried out to develop new frameworks for multi-class discrimination of malignant diseases.

Declaration of competing interest

PLEASE CHECK THE FOLLOWING AS APPROPRIATE.

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- The Article I have submitted to the journal for review is original, has been written by the stated authors and has not been published elsewhere.
- The Images that I have submitted to the journal for review are original, was taken by the stated authors, and has not been published elsewhere.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- The below authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Acknowledgements

This work was supported by the Swiss National Science Foundation under Grant No. 320030_176052.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104752.

References

[1] Z. Cai, Q. Liu, Understanding the global cancer statistics 2018: implications for cancer control, Sci. China Life Sci. (2019) 1–4.

[2] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. Parkin, M. Piñeros, A. Znaor, F. Bray, Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, Int. J. Canc. 144 (2019) 1941–1953.

[3] J.R. Molina, P. Yang, S.D. Cassivi, S.E. Schild, A.A. Adjei, Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship, Mayo Clin. Proc. 83 (2008) 584–594.

[4] S. Navada, P. Lai, A. Schwartz, G. Kalemkerian, Temporal trends in small cell lung cancer: analysis of the national Surveillance, Epidemiology, and End-Results (SEER) database, J. Clin. Oncol. 24 (2006), 7082-7082.

[5] N. Duma, R. Santana-Davila, J.R. Molina, Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment, Mayo Clin. Proc. 94 (2019) 1623–1640.

[6] W.D. Travis, E. Brambilla, A.G. Nicholson, Y. Yatabe, J.H. Austin, M.B. Beasley, L. R. Chirieac, S. Dacic, E. Duhig, D.B. Flieder, The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification, J. Thorac. Oncol. 10 (2015) 1243–1260.

[7] D.B. Doroshow, R.S. Herbst, Treatment of advanced non–small cell lung cancer in 2018, JAMA Oncol 4 (2018) 569–570.

[8] D.A. Dziedzic, P. Rudzinski, R. Langfort, T. Orlowski, Risk factors for local and distant recurrence after surgical treatment in patients with non–small-cell lung cancer, Clin. Lung Canc. 17 (2016) e157–e167.

[9] C. Biancosino, M. Krüger, E. Vollmer, L. Welker, Intraoperative fine needle aspirations-diagnosis and typing of lung cancer in small biopsies: challenges and limitations, Diagn. Pathol. 11 (2016) 59.

[10] P.B. Bach, J.R. Jett, U. Pastorino, M.S. Tockman, S.J. Swensen, C.B. Begg, Computed tomography screening and lung cancer outcomes, J. Am. Med. Assoc. 297 (2007) 953–961.

[11] B. Jiang, S. Takashima, C. Miyake, T. Hakucho, Y. Takahashi, D. Morimoto, H. Numasaki, K. Nakanishi, Y. Tomita, M. Higashiyama, Thin-section CT findings in peripheral lung cancer of 3 cm or smaller: are there any characteristic features for predicting tumor histology or do they depend only on tumor size? Acta Radiol. 55 (2014) 302–308.

[12] R. Kakinuma, K. Kodama, K. Yamada, A. Yokoyama, S. Adachi, K. Mori, Y. Fukuda, K. Kuriyama, J. Oda, M. Noguchi, Performance evaluation of 4 measuring methods of ground-glass opacities for predicting the 5-year relapse-free survival of patients with peripheral nonsmall cell lung cancer: a multicenter study, J. Comput. Assist. Tomogr. 32 (2008) 792–798.

[13] R. Patil, G. Mahadevaiah, A. Dekker, An approach toward automatic classification of tumor histopathology of non–small cell lung cancer based on radiomic features, Tomography 2 (2016) 374.

[14] W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, H.J. Aerts, Exploratory study to identify radiomics classifiers for lung cancer histology, Front Oncol 6 (2016) 71.

[15] X. Zhu, D. Dong, Z. Chen, M. Fang, L. Zhang, J. Song, D. Yu, Y. Zang, Z. Liu, J. Shi, Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer, Eur. Radiol. 28 (2018) 2772–2778.

[16] M. Nazari, I. Shiri, G. Hajianfar, N. Oveisi, H. Abdollahi, M.R. Deevband, M. Oveisi, H. Zaidi, Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning, Radiol. Med. (2020) 754–762.

[17] P. Lambin, R.T. Leijenaar, T.M. Deist, J. Peerlings, E.E. De Jong, J. Van Timmeren, S. Sanduleanu, R.T. Larue, A.J. Even, A. Jochems, Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (2017) 749–762.

[18] S.P. Shayesteh, M. Nazari, A. Salahshour, S. Sandoughdaran, G. Hajianfar, M. Khateri, A.Y. Joybari, F. Jozian, S.H.F. Feyzabad, H. Arabi, I. Shiri, H. Zaidi, Treatment response prediction using MRI-based pre-, post- and delta-radiomic features and machine learning algorithms in colorectal cancer, Med. Phys. 48 (7) (2021) 3691–3701, https://doi.org/10.1002/mp.14896.

[19] H. Arabi, A. AkhavanAllaf, A. Sanaat, I. Shiri, H. Zaidi, The promise of artificial intelligence and deep learning in PET and SPECT imaging, Phys. Med. 83 (2021) 122–137.

[20] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H. J.W.L. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, Radiomics: the process and the challenges, Magn. Reson. Imaging 30 (2012) 1234–1248.

[21] G. Hajianfar, I. Shiri, H. Maleki, N. Oveisi, A. Haghparast, H. Abdollahi, M. Oveisi, Noninvasive O6 methylguanine-DNA methyltransferase status prediction in glioblastoma multiforme cancer using magnetic resonance imaging radiomics features: univariate and multivariate radiogenomics analysis, World Neurosurg 132 (2019) e140–e161.

[22] I. Shiri, M. Sorouri, P. Geramifar, M. Nazari, M. Abdollahi, Y. Salimi, B. Khosravi, D. Askari, L. Aghaghazvini, G. Hajianfar, A. Kasaeian, H. Abdollahi, H. Arabi, A. Rahmim, A.R. Radmard, H. Zaidi, Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients, Comput. Biol. Med. 132 (2021) 104304.

[23] M. Nazari, I. Shiri, H. Zaidi, Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients, Comput. Biol. Med. 129 (2021) 104135.

[24] J.R.F. Junior, M. Koenigkam-Santos, F.E.G. Cipriano, A.T. Fabro, P.M. de Azevedo-Marques, Radiomics-based features for pattern recognition of lung cancer histopathology and metastases, Comput. Methods Progr. Biomed. 159 (2018) 23–30.

[25] Y. Han, Y. Ma, Z. Wu, F. Zhang, D. Zheng, X. Liu, L. Tao, Z. Liang, Z. Yang, X. Li, Histologic subtype classification of non-small cell lung cancer using PET/CT images, Eur. J. Nucl. Med. Mol. Imag. (2020) 1–11.

[26] J. Liu, J. Cui, F. Liu, Y. Yuan, F. Guo, G. Zhang, Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model, Med. Phys. 46 (2019) 3091–3100.

[27] J.H. Friedman, C.B. Roosen, An Introduction to Multivariate Adaptive Regression Splines, Sage Publications Sage CA, Thousand Oaks, CA, 1995.

[28] R.W. Kursa Mb, Feature selection with the Boruta package, J. Stat. Software 36 (11) (2010) 1–13.

[29] V.B. Clark K, K. Smith, et al., The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imag. 26 (6) (2013) 1045–1057.

[30] M. Hatt, M. Vallieres, D. Visvikis, A. Zwanenburg, IBSI: an international community radiomics standardization initiative, J. Nucl. Med. 59 (2018), 287-287.

[31] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillon-Robin, S. Pieper, H.J.W.L. Aerts, Computational radiomics system to decode the radiographic phenotype, Canc. Res. 77 (21) (2017) e104–e107.

[32] M. Shafiq-Ul-Hassan, K. Latifi, G. Zhang, G. Ullah, R. Gillies, E. Moros, Voxel size and gray level normalization of CT radiomic features in lung cancer, Sci. Rep. 8 (2018) 10545.

[33] M. Shafiq-Ul-Hassan, G.G. Zhang, K. Latifi, G. Ullah, D.C. Hunt, Y. Balagurunathan, M.A. Abdalah, M.B. Schabath, D.G. Goldgof, D. Mackin, L.E. Court, R.J. Gillies, E. G. Moros, Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels, Med. Phys. 44 (2017) 1050–1062.

[34] R.M. Haralick, Statistical and structural approaches to texture, Proc. IEEE 67 (1979) 786–804.

[35] H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, Corrigendum: decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (1) (2014) 4644.

[36] M.B. Kursa, W.R. Rudnicki, M.M.B. Kursa, Package 'Boruta', 2020.

[37] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, C.-J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Comput. Stat. Data Anal. 50 (2006) 1113–1130.

[38] L.J. Isaksson, M. Pepa, M. Zaffaroni, G. Marvaso, D. Alterio, S. Volpe, G. Corrao, M. Augugliaro, A. Starzyńska, M.C. Leonardi, Machine learning-based models for prediction of toxicity outcomes in radiotherapy, Front Oncol 10 (2020) 790.

[39] B. Ripley, W. Venables, M.B. Ripley, Package 'nnet', R package version 7 (2016) 3–12.

[40] D.L. Ball, R.J. Fisher, B.H. Burmeister, M.G. Poulsen, P.H. Graham, M. G. Penniment, S.K. Vinod, H.E. Krawitz, D.J. Joseph, G.C. Wheeler, The complex relationship between lung tumor volume and survival in patients with non-small cell lung cancer treated by definitive radiotherapy: a prospective, observational prognostic factor study of the Trans-Tasman Radiation Oncology Group (TROG 99.05), Radiother. Oncol. 106 (2013) 305–311.

[41] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. (1991) 1–67.

[42] V.L. Miguéis, A. Camanho, J.F. e Cunha, Customer attrition in retailing: an application of multivariate adaptive regression splines, Expert Syst 40 (2013) 6225–6232.

[43] M.B. Kursa, Robustness of Random Forest-based gene selection methods, BMC Bioinf. 15 (2014) 8.

[44] L.N. Sanchez-Pinto, L.R. Venable, J. Fahrenbach, M.M. Churpek, Comparison of variable selection methods for clinical predictive modeling, Int. J. Med. Inf. 116 (2018) 10–17.

[45] Y. Guo, Q. Song, M. Jiang, Y. Guo, P. Xu, Y. Zhang, C.C. Fu, Q. Fang, M. Zeng, X. Yao, Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics, Acad. Radiol. (2021) in press, https://doi.org/10.1016/j.acra.2020.06.010.

[46] I. El Naqa, D. Ruan, G. Valdes, A. Dekker, T. McNutt, Y. Ge, Q.J. Wu, J.H. Oh, M. Thor, W. Smith, A. Rao, C. Fuller, Y. Xiao, F. Manion, M. Schipper, C. Mayo, J. M. Moran, R. Ten Haken, Machine learning and modeling: data, validation, communication challenges, Med. Phys. 45 (2018) e834–e840.

[47] A. Haga, W. Takahashi, S. Aoki, K. Nawa, H. Yamashita, O. Abe, K. Nakagawa, Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis, Radiol Phys Technol 11 (2018) 27–35.

[48] U. Bashir, B. Kawa, M. Siddique, S.M. Mak, A. Nair, E. Mclean, A. Bille, V. Goh, G. Cook, Non-invasive classification of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features, Br. J. Radiol. 92 (2019), 20190159.

[49] K. Skogen, B. Ganeshan, C. Good, G. Critchley, K. Miles, Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade, J. Neuro Oncol. 111 (2013) 213–219.

[50] B. Ganeshan, V. Goh, H.C. Mandeville, Q.S. Ng, P.J. Hoskin, K.A. Miles, Non–small cell lung cancer: histopathologic correlates for texture parameters at CT, Radiology 266 (2013) 326–336.

[51] M.G. Lubner, N. Stabo, E.J. Abel, A.M. del Rio, P.J. Pickhardt, CT textural analysis of large primary renal cell carcinomas: pretreatment tumor heterogeneity correlates with histologic findings and clinical outcomes, AJR Am. J. Roentgenol. 207 (2016) 96–105.

[52] E. Linning, L. Lu, L. Li, H. Yang, L.H. Schwartz, B. Zhao, Radiomics for classification of lung cancer histological subtypes based on nonenhanced computed tomography, Acad. Radiol. 26 (2019) 1245–1252.
[53] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. 16 (2002) 321–357.
[54] S.M. Abd Elrahman, A. Abraham, A review of class imbalance problem, Journal of Network and Innovative Computing 1 (2013) 332–340.
[55] T.W. Way, B. Sahiner, L.M. Hadjiiski, H.P. Chan, Effect of finite sample size on feature selection and classification: a simulation study, Med. Phys. 37 (2010) 907–920.
[56] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem: a review, Int. J. Advance Soft Compu. Appl 7 (2015) 176–204.
[57] N. Japkowicz, Class Imbalances: Are We Focusing on the Right Issue, Workshop on Learning from Imbalanced Data Sets II, 2003, p. 63.
[58] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. 6 (2002) 429–449.
[59] S. Das, Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection, Icml, 2001, pp. 74–81.
[60] X. Tang, X. Xu, Z. Han, G. Bai, H. Wang, Y. Liu, P. Du, Z. Liang, J. Zhang, H. Lu, H. Yin, Elaboration of a multimodal MRI-based radiomics signature for the preoperative prediction of the histological subtype in patients with non-small-cell lung cancer, Biomed. Eng. Online 19 (2020) 5.
[61] M. El-Zein, M.E. Parent, B. Nicolau, A. Koushik, J. Siemiatycki, M.C. Rousseau, Body mass index, lifetime smoking intensity and lung cancer risk, Int. J. Canc. 133 (2013) 1721–1731.
[62] B. Zhao, Y. Tan, W.Y. Tsai, L.H. Schwartz, L. Lu, Exploring variability in CT characterization of tumors: a preliminary phantom study, Transl Oncol 7 (2014) 88–93.
[63] M. Edalat-Javid, I. Shiri, G. Hajianfar, H. Abdollahi, H. Arabi, N. Oveisi, M. Javadian, M. Shamsaei Zafarghandi, H. Malek, A. Bitarafan-Rajabi, Cardiac SPECT radiomic features repeatability and reproducibility: a multi-scanner phantom study, J. Nucl. Cardiol. (2021) in press.
[64] I. Shiri, G. Hajianfar, A. Sohrabi, H. Abdollahi, S.P. Shayesteh, P. Geramifar, H. Zaidi, M. Oveisi, A. Rahmim, Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: test–retest and image registration analyses, Med. Phys. 47 (2020) 4265–4280.
[65] C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R.H. Mak, S. Mitra, B.U. Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin, H.J. Aerts, Robust Radiomics feature quantification using semiautomatic volumetric segmentation, PloS One 9 (2014), e102107.
[66] L.A. Hunter, S. Krafft, F. Stingo, H. Choi, M.K. Martel, S.F. Kry, L.E. Court, High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images, Med. Phys. 40 (2013) 121916.
[67] B. Zhao, Y. Tan, W.Y. Tsai, J. Qi, C. Xie, L. Lu, L.H. Schwartz, Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (2016) 23428.