










Article

Machine Learning and Feature Selection Methods for *EGFR* Mutation Status Prediction in Lung Cancer

Joana Morgado ^{1,2,*} , Tania Pereira ¹ , Francisco Silva ¹ , Cláudia Freitas ^{3,4} , Eduardo Negrão ³,
Beatriz Flor de Lima ³, Miguel Correia da Silva ³ , António J. Madureira ^{3,4}, Isabel Ramos ^{3,4},
Venceslau Hespanhol ^{3,4} , José Luis Costa ^{4,5,6} , António Cunha ^{1,7}  and Hélder P. Oliveira ^{1,2} 

- ¹ INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal; tania.pereira@inesctec.pt (T.P.); francisco.c.silva@inesctec.pt (F.S.); antonio.cunha@inesctec.pt (A.C.); helder.f.oliveira@inesctec.pt (H.P.O.)
- ² FCUP—Faculty of Science, University of Porto, 4169-007 Porto, Portugal
- ³ CHUSJ—Centro Hospitalar e Universitário de São João, 4200-319 Porto, Portugal; claudiaasfreitas@gmail.com (C.F.); eduardo.negrao@gmail.com (E.N.); beatrizflordelima@hotmail.com (B.F.d.L.); miguel.ncds@gmail.com (M.C.d.S.); antonio.madureira@chs.min-saude.pt (A.J.M.); radiologia.hsj@gmail.com (I.R.); hespanholv@gmail.com (V.H.)
- ⁴ FMUP—Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal; jcosta@ipatimup.pt
- ⁵ i3S—Institute for Research and Innovation in Health of the University of Porto, 4200-135 Porto, Portugal
- ⁶ IPATIMUP—Institute of Molecular Pathology and Immunology of the University of Porto, 4200-135 Porto, Portugal
- ⁷ INESC UTAD—Institute for Systems and Computer Engineering, University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal
- * Correspondence: joana.p.morgado@inesctec.pt



Citation: Morgado, J.; Pereira, T.; Silva, F.; Freitas, C.; Negrão, E.; de Lima, B.F.; da Silva, M.C.; Madureira, A.J.; Ramos, I.; Hespanhol, V.; et al. Machine Learning and Feature Selection Methods for *EGFR* Mutation Status Prediction in Lung Cancer. *Appl. Sci.* **2021**, *11*, 3273. <https://doi.org/10.3390/app11073273>

Academic Editor: Donato Cascio

Received: 17 March 2021

Accepted: 29 March 2021

Published: 6 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The evolution of personalized medicine has changed the therapeutic strategy from classical chemotherapy and radiotherapy to a genetic modification targeted therapy, and although biopsy is the traditional method to genetically characterize lung cancer tumor, it is an invasive and painful procedure for the patient. Nodule image features extracted from computed tomography (CT) scans have been used to create machine learning models that predict gene mutation status in a noninvasive, fast, and easy-to-use manner. However, recent studies have shown that radiomic features extracted from an extended region of interest (ROI) beyond the tumor, might be more relevant to predict the mutation status in lung cancer, and consequently may be used to significantly decrease the mortality rate of patients battling this condition. In this work, we investigated the relation between image phenotypes and the mutation status of Epidermal Growth Factor Receptor (*EGFR*), the most frequently mutated gene in lung cancer with several approved targeted-therapies, using radiomic features extracted from the lung containing the nodule. A variety of linear, nonlinear, and ensemble predictive classification models, along with several feature selection methods, were used to classify the binary outcome of wild-type or mutant *EGFR* mutation status. The results show that a comprehensive approach using a ROI that included the lung with nodule can capture relevant information and successfully predict the *EGFR* mutation status with increased performance compared to local nodule analyses. Linear Support Vector Machine, Elastic Net, and Logistic Regression, combined with the Principal Component Analysis feature selection method implemented with 70% of variance in the feature set, were the best-performing classifiers, reaching Area Under the Curve (AUC) values ranging from 0.725 to 0.737. This approach that exploits a holistic analysis indicates that information from more extensive regions of the lung containing the nodule allows a more complete lung cancer characterization and should be considered in future radiogenomic studies.

Keywords: radiogenomics; machine learning; feature selection; lung cancer; *EGFR* prediction

1. Introduction

Lung cancer is the foremost determinant cancer death amongst men and women, killing a vaster number of people than colon, breast, and prostate cancers combined [1]. This is linked to the fact that it is often diagnosed in an advanced stage, with 5% or less chance of a 5-year survival [2]. Non-Small-Cell Lung Carcinoma (NSCLC) is the most prevalent histological type of lung cancer, covering about 85% of all lung cancer cases [3], and Epidermal Growth Factor Receptor (*EGFR*) is the most frequently mutated gene that springs lung cancer of type Adenocarcinoma [4]. The cell surface receptor *EGFR* is responsible for cell growth and survival and its mutations promote *EGFR* permanent activation, which contributes to uncontrolled cell division [5,6]. This genomic biomarker with clinically approved therapies is now considered a strong prognostic indicator in lung cancer, rising opportunities to explore treatment strategies that rely on the individual's genetic profile [7]. Currently, biopsy is the primary method for characterizing lung cancer and identifying *EGFR* mutation status, using an extracted tumor tissue sample for molecular analysis. Nevertheless, this invasive procedure can lead to some associated side effects, as it can be painful and risky for the patient. Recently, blood-based screening has been used to detect early lung cancer diagnostic biomarkers [8]. However, despite being less invasive than biopsy, this technique is still bothersome for the patient.

Computed tomography (CT) scans provide a reliable lung cancer characterization, offering a faster and less invasive approach compared to traditional tissue biopsy [9]. Thereby, foretelling gene mutation status by CT can help determine the most appropriate treatment for each subject, while decreasing medical complications [10]. Extracting quantitative features from CT images that are then used as inputs within a predictive model that can directly classify the *EGFR* mutation status for lung cancer patients composes the essence of radiogenomics, a field that correlates the radiomic features (image phenotype) and genetic information (genotype) [11].

There are a few studies that used traditional statistical analysis and machine learning (ML)-based approaches to demonstrate that the *EGFR* mutation status is correlated with CT scan imaging phenotypes. Regarding the works that employed logistic regression models, Digumarthy et al. [12] obtained an Area Under the Curve (AUC) of 0.73 applying only radiomic features, which increased to 0.79 when clinical data was added. The study by Mei et al. [13] resulted in an AUC of 0.58 and 0.66 implementing only radiomic features and combination of radiomic and clinical features, respectively. Similarly, Liu et al. [14] showed that adding radiomic features to a clinical model resulted in a significant improvement of predicting power, as the AUC increased from 0.67 to 0.71, and Liu et al. [15] showed that using clinical variables combined with CT features (AUC = 0.78) resulted in higher AUC values, compared to using clinical variables alone (AUC = 0.69). Concerning the ML domain, a decision tree was built to predict the presence of *EGFR* mutations using a combination of four image features (emphysema, airway abnormality, the percentage of ground glass component and the type of tumor margin), resulting in a test set performance of 0.89 AUC [5]. A work based on a deep learning technique with automatic nodule radiomic feature-learning ability achieved an AUC of 0.69 and showed a significant improvement when hand-crafted CT features were combined with clinical characteristics (AUC = 0.81) [10]. Velazquez et al. [16] found a radiomic signature related to radiographic heterogeneity that successfully discriminated between mutant and wild-type *EGFR* cases (AUC = 0.69). Combining this signature with a clinical model of *EGFR* status (AUC = 0.70) significantly improved prediction performance (AUC = 0.75). A XGBoost classifier using only radiomic and semantic features from the nodule obtained an AUC of 0.58 and 0.65, respectively [4].

Previous studies have suggested that combining radiomic and semantic data allows the development of integrated predictors that exhibit improved performance compared to learning models that use only one type of data. However, there is an urgent need to create learning models that use only radiomic features that are automatically extracted from images, as these features are more objective than semantic data and allow to reduce the

dependency on data annotation by experts in assessing mutational status while reducing human error. Furthermore, the automatic detection of radiomic features from CT images is an important tool to help radiologists with the extensive and exhaustive work of CT annotation and overcome the lack of mass annotated datasets. Generally, the nodule is the main focus for lung cancer malignancy assessment and follow-up based on the well-established Fleischner Society and Lung-RADs Guidelines [17,18], and previous studies have analyzed only this important cluster of tumor cells for *EGFR* mutation status prediction [10,14,16]. Exploratory studies have recently shown that there is a correlation of *EGFR* mutation status with other lung diseases, such as emphysema and fibrosis, which seems to indicate that cancer development is related to multiple physiological changes not restricted to the nodule region [19,20], and that the inclusion of extratumor features allows a significant increase in *EGFR* mutation predictive performance [4,5,21].

Therefore, the current work applies a comprehensive approach for predicting *EGFR* mutation status, using only objective radiomic features directly extracted from a region of interest (ROI) containing the entire lung where the nodule is located. The hypothesis that it might be possible to develop predictive models with enhanced performance by combining nodule-related features with features from other lung structures is the main motivation for this work. This study includes combinations of six ML classification models—Logistic Regression, Elastic Net, Support Vector Machine (SVM) with linear and radial basis function (RBF) kernels, Random Forest, and Extreme Gradient Boosting (XGBoost)—and nine feature selection methods—Pairwise correlation filter, Principal component analysis (PCA) with feature set variance ranging from 65% to 95%, and QR decomposition, along with a baseline where no feature selection is used.

2. Materials and Methods

This section presents the dataset used in the current study, the feature extraction techniques and the classification methods employed. The pipeline implemented to predict the *EGFR* mutation status is represented in Figure 1.

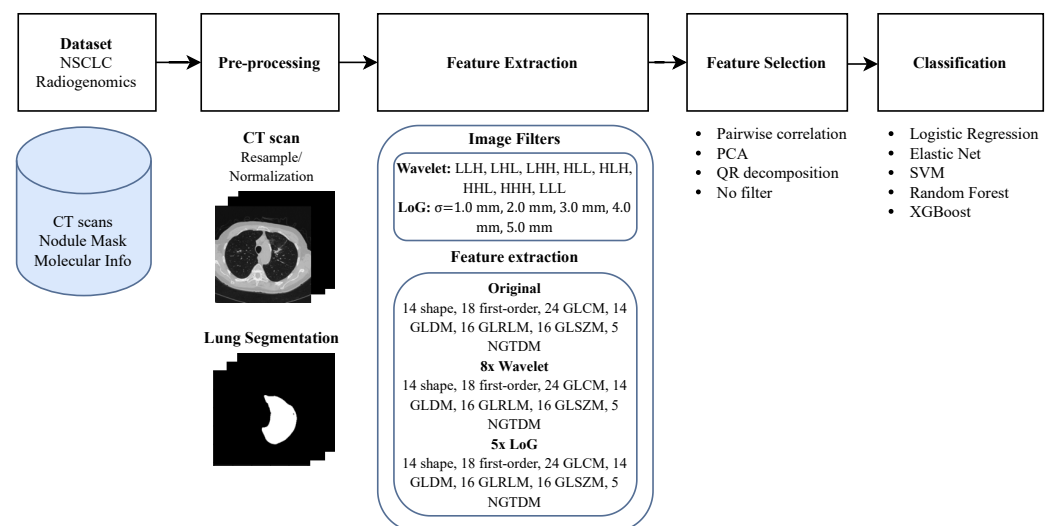


Figure 1. Overview of the classification approach for Epidermal Growth Factor Receptor (*EGFR*) mutation status prediction. Computed tomography (CT) images and segmentation masks of the lung are loaded into the software. Then, the wavelet and Laplacian of Gaussian (LoG) filters are applied to the original image and 1316 radiomic features of the region of interest (ROI) are extracted using the *Pyradiomics* package [22]. This process is performed for all images in the Non-Small-Cell Lung Carcinoma (NSCLC) dataset. Finally, various combinations of six classification models, three feature selection techniques, and a baseline without feature selection are implemented and compared.

2.1. Dataset

In order to fulfill the objectives set for this work, the NSCLC-Radiogenomics Dataset [23] was identified and used, as it comprises three fundamental requirements that were crucial for the development of the present study: thoracic CT scans from NSCLC patients, tumor binary segmentations, and *EGFR* mutation status labels. This dataset is a publicly available collection of CT images from a NSCLC cohort of 211 subjects collected retrospectively between 2008 and 2012 at Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System. The CT scans are paired with patient clinical history and were obtained using different scanner models and scanning protocols, presenting variations in slice thickness from 0.625 to 3 mm (median: 1.5 mm) and X-ray tube current from 124 to 699 mA (mean: 220 mA) at 80–140 kVp (mean: 120 kVp). Additionally, binary nodule segmentation masks are also stored in this database, as well as semantic tumor annotations. This dataset is the only public dataset that comprises paired information on lung-cancer-related gene mutation status and CT data; however, the present study focuses only on *EGFR* due to its clinical relevance based on approved target therapies. Out of 211 subjects, only 116 lung cancer patients from this database were considered since only these owned tumor binary masks and a *EGFR* mutation test result (mutant: 20%, wild-type: 80% (see Figure 2)).

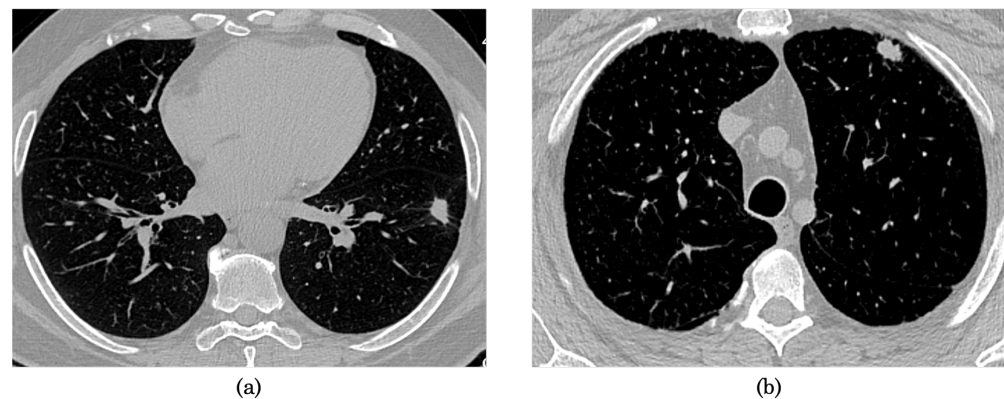


Figure 2. Representation of lung axial slices with nodule from the NSCLC-Radiogenomics dataset of patients with (a) mutant *EGFR* and (b) wild-type *EGFR*.

2.2. Pre-Processing

Firstly, the CT image pixel values were converted from radiodensity values to Hounsfield Units (HU) using the Rescale Slope and Rescale Intercept attributes stored in the metadata of the scans. Then, the entire dataset (including the tumor masks) was resampled to standardize image representations. The space between slices and pixel spacing were set to 1 mm and [1.0, 1.0] mm, respectively, and each slice dimension was calculated to match this new spacing, obtaining the resampled image by interpolation [24]. Additionally, all CT images were normalized between -1000 HU and 400 HU using the *min-max* normalization method. CT values below -1000 HU, corresponding to the radiodensity of air, were set to 0; values above 400 HU, representing hard tissues, were fixed to 1. A linear transformation was computed to map all the intermediate HU values to the $[0,1]$ range. Lung binary masks were segmented using a 2D lung segmentation model based on the U-Net architecture [25]. These lung regions underwent the same resampling operation to match the dimensions of the corresponding CT, and all images were rescaled to a size of $N \times 256 \times 256$ pixels, with N representing the number of slices of the correspondent CT scan. To obtain the CT images with only the lung containing the nodule, the mask of the nodule was superimposed on the lung segmentation in order to identify the lung where the nodule is located in the selected axial CT slice. Then, the image was transformed to consider half of the opposite side as background.

2.3. Feature Extraction

The 1316 radiomic features available for this study quantified lung characteristics from CT images and were extracted using the open source package *Pyradiomics* [22]. Lung voxels were used in the extraction of seven classes of features: shape-based (14 features), first-order (18 features), Gray Level Co-occurrence Matrix (GLCM) (24 features), Gray Level Dependence Matrix (GLDM) (14 features), Gray Level Run Length Matrix (GLRLM) (16 features), Gray Level Size Zone Matrix (GLSZM) (16 features) and Neighboring Gray Tone Difference Matrix (NGTDM) (5 features). Shape features examine the size and shape of the ROI, employing only the lung segmentation masks in the calculations. First-order features describe the distribution of voxel intensities within the ROI using basic metrics such as mean, median, range, and standard deviation. GLCM features describe the second-order joint probability function of the ROI, while GLDM features quantify gray level dependencies in the image. GLRLM features quantify the length of consecutive pixels that have the same gray level intensity and GLSZM features quantify the number of connected voxels that share the same gray level value. Finally, NGTDM features characterize the difference between the gray value of a pixel and the average gray value of the neighboring pixels within a defined distance [26].

The radiomic features were computed both on the original image (107 features) and on images obtained after application of wavelet (744 features) and Laplacian of Gaussian (LoG) (465 features) filters. The wavelet transform applies a wavelet filter to each CT image, which is then decomposed in low and high frequencies into 8 different images. It applies either a high-pass filter (represented as H) or a low-pass filter (represented as L) in each one of the x, y, and z directions: LLH, LHL, LHH, HLL, HLH, HHL, HHH and LLL [27]. The LoG filter yields a derived image for each applied sigma value in order to emphasize areas of gray level change, where sigma defines how coarse the emphasized texture should be [28]. In these studies, five filters with different sigma values were applied (sigma = 1.0 mm, 2.0 mm, 3.0 mm, 4.0 mm, 5.0 mm).

2.4. Feature Selection Methods

The radiomic features were subjected to a feature selection process in order to prevent overfitting, improve learning accuracy, and reduce computation time [29]. To this end, we considered three feature selection techniques widely used in the radiogenomics field: pairwise correlation filter, PCA, and QR decomposition. Additionally, a baseline was also implemented where no feature selection method was used to compare the results obtained with and without feature selection.

The pairwise correlation filter removes variables whose pairwise correlation is greater than a specific cutoff. First, a correlation matrix is created with values representing the pairwise correlations for all feature combinations. Then, features that have an absolute pairwise correlation equal to or greater than the cutoff are excluded [30]. After investigating multiple cutoffs, the cutoff value was set to 0.95.

PCA is a method that projects high-dimensional data into a new lower dimensional representation while keeping all relevant linear structure intact. This method generates new, uncorrelated variables that explain a large proportion of the variance in the original feature space [31]. PCA was implemented at seven different cutoffs, where the number of components accounted for percentages of variance in the feature set ranging from 65% to 95%.

The QR decomposition along with an iterative procedure is used to remove features that are linear combinations of others. The feature matrix is decomposed into two matrices: the orthogonal matrix Q and the upper-triangular matrix R. The latter is used to determine which features are linearly dependent so that they can be sequentially removed [32].

2.5. Learning Models

Since we intended to explore different types of ML models, the predictive classifying models used in this study are from three different families: linear, nonlinear, and ensemble. The linear classifiers used include the Logistic Regression, Elastic Net, and Linear SVM

models, while the nonlinear classifier includes the SVM algorithm with an RBF kernel. Of the ensemble models, the Random Forest and the XGBoost were implemented.

Logistic Regression is a classification model that employs the sigmoid function as a cost function in order to return a probability value that can be mapped to discrete classes [33]. It is one of the simplest ML algorithms and is easy to implement, interpret, and very efficient to train. Elastic Net regression is a penalized linear regression model that imposes a linear combination of regularization penalties to the loss function during training. These norm regularizations include both the L1 and L2 penalties, which are based on the sum of the absolute coefficient values and the sum of the squared coefficient values, respectively [34].

SVM is a discriminative classifier that transforms the original feature space into a higher-dimensional space based on a user-defined kernel function and then finds support vectors to maximize the margin separating the classes. New unlabeled samples are classified according to the side of the hyperplane they lie on [35]. Linear SVMs can only find a decision boundary to classify linearly separable features. On the other hand, when the dataset is separable by a nonlinear boundary, SVM uses nonlinear kernel functions, such as the RBF, to overcome the curse of dimensionality and properly transform the feature space. SVMs are powerful yet flexible supervised ML algorithms that are used in a variety of applications, such as the diagnosis and prognosis of cancer and other diseases [36–40].

Random Forest classifier is a bagging-type ensemble method of decision trees that trains several trees in parallel and aggregates the decisions of individual trees to reach the final prediction. This classifier tends to outperform most other classification methods in terms of accuracy, variance, and bias, without overfitting issues [41]. XGBoost is a decision-tree-based ensemble algorithm that uses a gradient boosting framework and has been widely used in lung cancer studies [4,42–44]. This classifier has been shown to yield superior predictive results using less computing resources in the shortest amount of time compared to other models due to its parallel processing, tree-pruning, sparse data handling, and regularization to prevent overfitting [45].

2.6. Training

Data was randomly split into a training set (80%) and a test set (20%). The training and testing processes were repeated for 50 random splits of the original dataset to investigate data variance and for better performance robustness.

In the NSCLC-Radiogenomic dataset, the *EGFR* mutant is under-represented, resulting in classifiers with poor predictive performance for this minority class. To overcome the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied for each fold. This oversampling approach creates new random synthetic minority class instances between the lines that connect each one of the k nearest neighbors of each minority class sample [46]. This data augmentation technique made it possible to obtain a class-balanced training set (consisting of the same number of mutant and wild-type samples), which was then used to train the classifiers.

The classifier's hyperparameters were tuned through a 5-fold cross-validation randomized search on the training data. The range of hyperparameters considered for the six classifiers are presented in Tables S1–S6 of the supplementary material.

Since Elastic Net is a regression model and not a classification model, predictions for this algorithm were performed considering a decision threshold equal to 0.5. This means that normalized predicted probabilities less than 0.5 were assigned to the wild-type class *EGFR* and probability values greater than or equal to 0.5 were mapped to the mutant class *EGFR*. On the other hand, for the classification models, the optimal threshold was calculated from the Receiver Operating Characteristic (ROC) curve using the threshold-moving method. This method uses the original training set to train the model and then moves the decision threshold such that the minority class examples can be more easily correctly predicted.

2.7. Performance Metrics

For a better understanding of the predictive performance of each classifier and feature selection method, different evaluation metrics were averaged over the 50 random train-test splits. The AUC was computed, along with precision, sensitivity, and specificity.

3. Results

The number of radiomic features removed and retained after the implementation of the nine feature selection methods is shown in Table 1.

The hyperparameters that achieved the best performance on the test set for each classifier and feature selection method are shown in Tables S7–S12 of the supplementary material. Furthermore, in the supplementary material, the AUC, precision, sensitivity, and specificity metrics, for each classifier-feature selection method combination, can be found in Tables S13–S18.

Table 1. Number of removed and retained features after the implementation of the pairwise correlation filter, Principal component analysis (PCA) with feature set variance ranging from 65% to 95%, and QR decomposition.

Feature Selection Method	Number of Removed Features	Number of Retained Features
Pairwise correlation filter	1223	93
PCA 65%	1312	4
PCA 70%	1311	5
PCA 75%	1310	6
PCA 80%	1309	7
PCA 85%	1306	10
PCA 90%	1301	15
PCA 95%	1290	26
QR decomposition	988	328

Figure 3 gives the mean AUC (average value of all 50 results) for each classifier across the various feature selection methods in a heatmap form, whereas Table 2 summarizes the best performance results obtained in the present study for each one of the six classifiers. SVM with linear kernel, Elastic Net, and Logistic Regression classifiers had the best overall predictive performance and showed the best results when combined with the PCA feature selection method with 70% of variance, presenting AUC values greater than 0.7. It can thus be concluded that these classifiers performed well in predicting the *EGFR* mutation status. In fact, the highest AUC was obtained with the Linear SVM classifier and the PCA 70% (AUC = 0.737 ± 0.018). Furthermore, the implementation of the Logistic Regression classifier with PCA 70% resulted in the highest precision (Precision = 0.682 ± 0.099) and sensitivity (Sensitivity = 0.699 ± 0.039) values. On the other hand, the XGBoost classifier with PCA 70% achieved the highest specificity result (Specificity = 0.767 ± 0.017). Among all feature selection methods, PCA yielded the highest AUC results on average in five of the six classifiers studied, whereas QR decomposition held lower AUC values. The standard deviations of all results are fairly similar, which shows that no metric had a significantly higher or lower variance than the others.

Table 2. Performance results of each classifier with the best AUC. For each combination of classifier and feature selection method, the evaluation metrics AUC, precision, sensitivity, and specificity are presented as mean \pm standard deviation. AUC—Area Under the Curve; SVM—Support Vector Machine; XGBoost—Extreme Gradient Boosting; RBF—Radial Basis Function.

Classifier	Feature Selection	AUC	Precision	Sensitivity	Specificity
SVM (linear kernel)	PCA 70%	0.737 \pm 0.018	0.644 \pm 0.012	0.615 \pm 0.010	0.685 \pm 0.095
Elastic Net	PCA 70%	0.733 \pm 0.011	0.585 \pm 0.048	0.611 \pm 0.033	0.715 \pm 0.013
Logistic Regression	PCA 70%	0.725 \pm 0.012	0.682 \pm 0.099	0.699 \pm 0.039	0.743 \pm 0.079
XGBoost	PCA 70%	0.697 \pm 0.032	0.640 \pm 0.036	0.632 \pm 0.040	0.767 \pm 0.017
Random Forest	No filter	0.696 \pm 0.011	0.683 \pm 0.090	0.688 \pm 0.029	0.721 \pm 0.012
SVM (RBF kernel)	PCA 75%	0.583 \pm 0.014	0.601 \pm 0.008	0.530 \pm 0.017	0.623 \pm 0.009

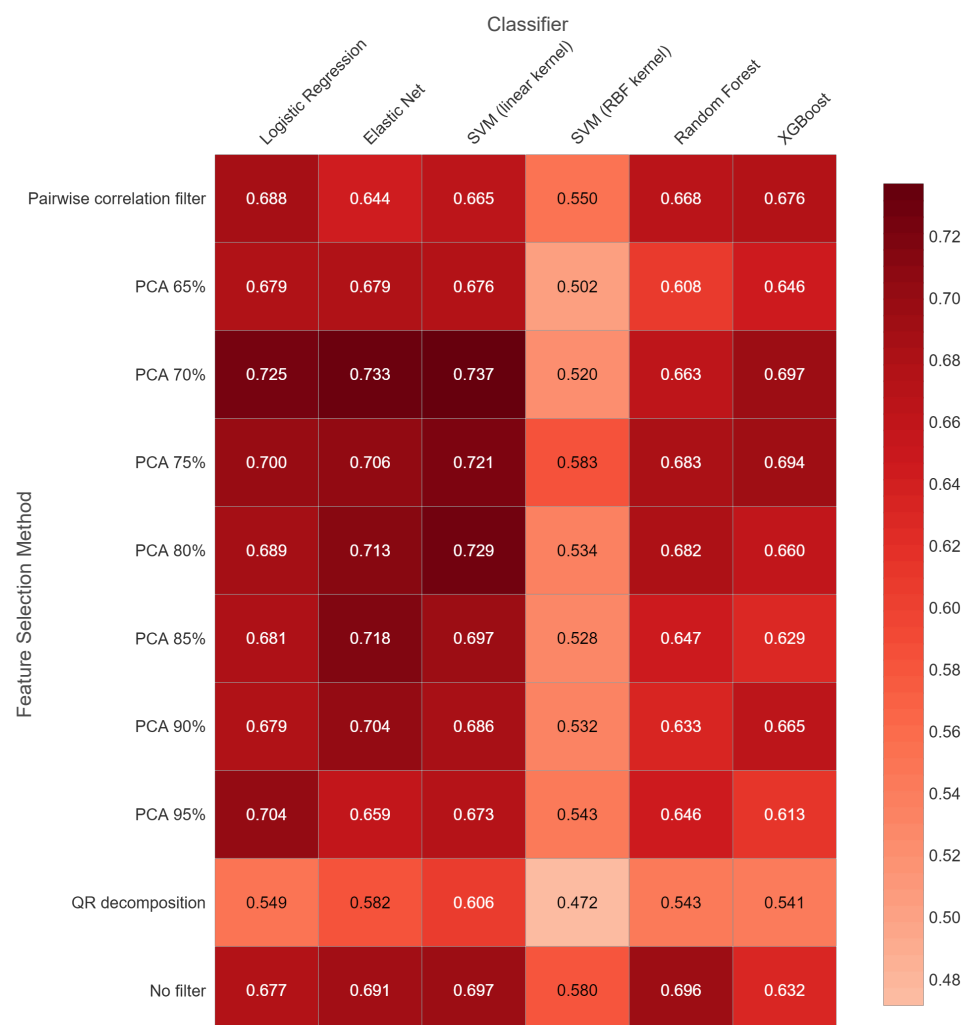


Figure 3. Heatmap with the AUC of each classifier/feature selection combination. Dark colors stand for the best results, while light colors represent the worst outcomes.

4. Discussion

In this study, we assessed *EGFR* mutation status, using radiomic features extracted from the entire lung volume containing the nodule on CT images. Six ML algorithms were trained and used in conjunction with pairwise correlation, PCA, and QR decomposition feature selection methods. These filtering techniques were employed since using too many features in the classification algorithm can lead to overfitting, in which noise or irrelevant features may exert undue influence on classification decisions. In addition, a baseline without feature selection method was also implemented to compare results obtained with and without feature selection.

The results of this study suggest that the linear learning models—SVM with linear kernel, Elastic Net, and Logistic Regression—perform well with quantitative imaging features as their predictors, whereas the SVM classifier based on the RBF kernel performs poorly. Additionally, the commonly used classifiers, Random Forest and XGBoost, show acceptable performance results. Furthermore, we found that the model less used in radiomics and radiogenomics, but commonly used in genomics—Elastic Net—was one of the best performing classifiers, as shown in a previous study that investigated the ability of various ML classifiers to accurately predict lung cancer nodule status [47]. The present work also indicates that the Linear SVM classifier in conjunction with PCA feature selection with 70% variance should be considered in future *EGFR* mutation status classification studies. In our opinion, it is crucial to highlight the fact that linear models obtained the best results in the present study, as it may change the direction and extend the use of simpler and more interpretable models in seemingly complex problems.

We also show that, in general, feature selection methods that reduce the number of features prior to model training appear to improve predictive performance compared to previous radiomic analyses [47–50]. In the radiogenomics field, a large number of features tend not to provide additional information because they are highly correlated and are linear combinations of others. However, in ensemble learning models, such as Random Forest and XGBoost, which perform automatic feature selection, additional feature filtering does not seem to have a significant impact on model performance. Considering the feature selection methods, it was not possible to identify the one that gives better results for all classifiers since the models are defined with different mathematical principles and optimize different parameters; as a consequence, different classifiers provide better results when combined with different numbers of radiomic features. Nevertheless, based on the best results per feature selection method, we recommend considering the PCA method with 70% and 75% of variance over the pairwise correlation filter and QR decomposition, as these two methods seem to remove very important features for model predictions.

Considering approaches that employ only radiomic features, the outcomes of this work show that radiomic characteristics from the entire lung containing the nodule provided better results in *EGFR* mutation status assessment compared to traditional nodule-based approaches and to other methodologies that consider other lung structures [4,5,14,51]. However, direct performance comparisons between models trained and tested with data from different datasets would not bring a fair discussion point to this study. Nevertheless, the results of the current study indicate that the CT features with the highest correlation with *EGFR* mutation are from the lung that has the nodule and these are therefore the main contributors to the model decision. It is crucial to highlight these results and further investigate the importance of holistic lung cancer characterization studies, as there are many complex combinations of morphological, molecular, and genetic alterations that occur during lung cancer development that, when taken into account, would allow the development of more accurate classifiers for *EGFR* mutation status prediction [21].

The biggest limitation of this work is the reduced size of the used dataset, which is unrepresentative of the population characteristics, making it difficult to find a relevant pattern for such a complex problem. Furthermore, deep learning approaches, such as convolutional neural networks (e.g., CNN and 2D-CNN) and recurrent neural networks (e.g., LSTM and BiLSTM), are powerful methods for *EGFR* mutation status assessment.

However, data limitation does not allow the use of these end-to-end techniques based on deep learning. To increase the chances of better *EGFR* mutation status predictions, it is necessary to develop reproducible, clinically viable, and robust predictive models that can handle population heterogeneities. For this, a large and heterogeneous cohort of patients affected by lung cancer is crucial, as well as methods capable of coping with data heterogeneity. However, data access and uniform acquisition, along with privacy issues, fees, and data management requirements are the main limitations in clinical and imaging data access. Another limitation relies on the fact that only the most frequent oncogene in lung cancer was studied. As future work, it is important to analyze other feature selection techniques, such as Linear discriminant analysis, which, in addition to being a prediction model, can be used as a dimensionality reduction technique and autoencoder-based architecture. Additionally, it is also important to analyze whether the feature selection methods implemented in the studies can be used, for example, as information gain or mutual information in order to better understand how the selection techniques affect the features and consequently the performance of the learning models. It is also relevant in the future to consider other lung-cancer-related genes in order to obtain a more complete characterization, which would have an important impact on new targeted personalized therapies.

5. Conclusions

Predicting *EGFR* mutation status by CT imaging can improve the determination of the most appropriate treatment for each lung cancer patient and is a less invasive alternative compared to the traditional biopsy. This study proposed a comprehensive approach for the classification of *EGFR* mutation status using only radiomic features extracted from a ROI containing the entire lung where the nodule is located, changing the direction of traditional approaches, which until now have been mainly focused on the nodule. The results obtained with this novel and holistic approach showed that information from more extensive regions of the lung containing the nodule allows for a more complete lung cancer characterization. Our work suggested that Linear SVM, Elastic Net, and Logistic Regression are the most robust models for *EGFR* mutation status prediction and supports their use by others in future radiogenomics studies. Furthermore, we recommend the application of methods that reduce the number of features prior to model training, in particular, PCA methods, as they seem to improve predictive performance. We also encourage the study and comparison of various features and modeling approaches for predicting *EGFR* mutation status since improvements in prediction are often achieved when different combinations are utilized.

Supplementary Materials: The following are available online at <https://www.mdpi.com/2076-3417/11/7/3273/s1>.

Author Contributions: J.M., T.P., F.S., A.C., and H.P.O. conceived the study; C.F. and V.H. provided the pneumology insights; E.N., B.F.d.L., M.C.d.S., I.R., and A.M. gave the radiology insights; and J.L.C. gave the molecular biology insights. A.J.M. and T.P. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by the ERDF—European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation—COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-030263.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data was obtained from the open-access NSCLC-Radiogenomics dataset publicly available at The Cancer Imaging Archive (TCIA) database [23,52]. Imaging and clinical data have been de-identified by TCIA and approved by the Institutional Review Board of the TCIA hosting institution. Ethical approval was reviewed and approved by the Washington

University Institutional Review Board protocols. Informed consent was obtained from all individual participants included in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [\[CrossRef\]](#)
2. Janssen-Heijnen, M.L.; Coebergh, J.W.W. Trends in incidence and prognosis of the histological subtypes of lung cancer in North America, Australia, New Zealand and Europe. *Lung Cancer* **2001**, *31*, 123–137. [\[CrossRef\]](#)
3. Molina, J.R.; Yang, P.; Cassivi, S.D.; Schild, S.E.; Adjei, A.A. Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. In *Mayo Clinic Proceedings*; Elsevier: Amsterdam, The Netherlands, 2008; Volume 83, pp. 584–594.
4. Pinheiro, G.; Pereira, T.; Dias, C.; Freitas, C.; Hespanhol, V.; Costa, J.L.; Cunha, A.; Oliveira, H.P. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Sci. Rep.* **2020**, *10*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Gevaert, O.; Echegaray, S.; Khuong, A.; Hoang, C.D.; Shrager, J.B.; Jensen, K.C.; Berry, G.J.; Guo, H.H.; Lau, C.; Plevritis, S.K.; et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Sci. Rep.* **2017**, *7*, 1–8. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Purba, E.R.; Saita, E.i.; Maruyama, I.N. Activation of the EGF receptor by ligand binding and oncogenic mutations: The “rotation model”. *Cells* **2017**, *6*, 13. [\[CrossRef\]](#)
7. Jiang, W.; Cai, G.; Hu, P.C.; Wang, Y. Personalized medicine in non-small cell lung cancer: A review from a pharmacogenomics perspective. *Acta Pharm. Sin. B* **2018**, *8*, 530–538. [\[CrossRef\]](#)
8. Wang, Y.; Liu, S.; Wang, Z.; Fan, Y.; Huang, J.; Huang, L.; Li, Z.; Li, X.; Jin, M.; Yu, Q.; et al. A Machine Learning-Based Investigation of Gender-Specific Prognosis of Lung Cancers. *Medicina* **2021**, *57*, 99. [\[CrossRef\]](#)
9. Ostridge, K.; Wilkinson, T.M. Present and future utility of computed tomography scanning in the assessment and management of COPD. *Eur. Respir. J.* **2016**, *48*, 216–228. [\[CrossRef\]](#)
10. Wang, S.; Shi, J.; Ye, Z.; Dong, D.; Yu, D.; Zhou, M.; Liu, Y.; Gevaert, O.; Wang, K.; Zhu, Y.; et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.* **2019**, *53*, 1–9. [\[CrossRef\]](#)
11. Bodalal, Z.; Trebeschi, S.; Nguyen-Kim, T.D.L.; Schats, W.; Beets-Tan, R. Radiogenomics: Bridging imaging and genomics. *Abdom. Radiol.* **2019**, *44*, 1960–1984. [\[CrossRef\]](#)
12. Digumarthy, S.R.; Padole, A.M.; Gullo, R.L.; Sequist, L.V.; Kalra, M.K. Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status? *Medicine* **2019**, *98*, 1–8. [\[CrossRef\]](#)
13. Mei, D.; Luo, Y.; Wang, Y.; Gong, J. CT texture analysis of lung adenocarcinoma: Can Radiomic features be surrogate biomarkers for EGFR mutation statuses. *Cancer Imaging* **2018**, *18*, 1–9. [\[CrossRef\]](#)
14. Liu, Y.; Kim, J.; Balagurunathan, Y.; Li, Q.; Garcia, A.L.; Stringfield, O.; Ye, Z.; Gillies, R.J. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clin. Lung Cancer* **2016**, *17*, 441–448. [\[CrossRef\]](#)
15. Liu, Y.; Kim, J.; Qu, F.; Liu, S.; Wang, H.; Balagurunathan, Y.; Ye, Z.; Gillies, R.J. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. *Radiology* **2016**, *280*, 271–280. [\[CrossRef\]](#)
16. Velazquez, E.R.; Parmar, C.; Liu, Y.; Coroller, T.P.; Cruz, G.; Stringfield, O.; Ye, Z.; Makrigiorgos, M.; Fennessy, F.; Mak, R.H.; et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Res.* **2017**, *77*, 3922–3930. [\[CrossRef\]](#)
17. MacMahon, H.; Naidich, D.P.; Goo, J.M.; Lee, K.S.; Leung, A.N.; Mayo, J.R.; Mehta, A.C.; Ohno, Y.; Powell, C.A.; Prokop, M.; et al. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017. *Radiology* **2017**, *284*, 228–243. [\[CrossRef\]](#)
18. Martin, M.D.; Kanne, J.P.; Broderick, L.S.; Kazerooni, E.A.; Meyer, C.A. Lung-RADS: Pushing the limits. *Radiographics* **2017**, *37*, 1975–1993. [\[CrossRef\]](#)
19. Dias, C.; Pinheiro, G.; Cunha, A.; Oliveira, H.P. Radiogenomics: Lung Cancer-Related Genes Mutation Status Prediction. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, Madrid, Spain, 1–4 July 2019; Springer: New York, NY, USA, 2019; pp. 335–345.
20. Zhang, H.; Cai, W.; Wang, Y.; Liao, M.; Tian, S. CT and clinical characteristics that predict risk of EGFR mutation in non-small cell lung cancer: A systematic review and meta-analysis. *Int. J. Clin. Oncol.* **2019**, *24*, 649–659. [\[CrossRef\]](#)
21. Pereira, T.; Freitas, C.; Costa, J.L.; Morgado, J.; Silva, F.; Negrão, E.; de Lima, B.F.; da Silva, M.C.; Madureira, A.J.; Ramos, I.; et al. Comprehensive Perspective for Lung Cancer Characterisation Based on AI Solutions Using CT Images. *J. Clin. Med.* **2021**, *10*, 118. [\[CrossRef\]](#)
22. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [\[CrossRef\]](#)
23. Bakr, S.; Gevaert, O.; Echegaray, S.; Ayers, K.; Zhou, M.; Shafiq, M.; Zheng, H.; Benson, J.A.; Zhang, W.; Leung, A.N.; et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **2018**, *5*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Limkin, E.J.; Reuzé, S.; Carré, A.; Sun, R.; Schernberg, A.; Alexis, A.; Deutsch, E.; Ferté, C.; Robert, C. The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. *Sci. Rep.* **2019**, *9*, 1–12. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Silva, F.; Pereira, T.; Frade, J.; Mendes, J.; Freitas, C.; Hespanhol, V.; Costa, J.L.; Cunha, A.; Oliveira, H.P. The Impact of Interstitial Diseases Patterns on Lung CT Segmentation. In *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Guadalajara, Mexico, 31 October–4 November 2021.

26. Meijer, K. Accuracy and Stability of Radiomic Features for Characterising Tumour Heterogeneity Using Multimodality Imaging: A Phantom Study. Master's Thesis, University of Twente, Twente, The Netherlands, 2019.
27. Procházka, A.; Gráfová, L.; Vyšata, O.; Caregroup, N. Three-dimensional wavelet transform in multi-dimensional biomedical volume processing. In Proceedings of the of the IASTED International Conference on Graphics and Virtual Reality, Cambridge, UK, 11–13 July 2011; Volume 263, p. 268.
28. Fotin, S.V.; Reeves, A.P.; Biancardi, A.M.; Yankelevitz, D.F.; Henschke, C.I. A multiscale Laplacian of Gaussian filtering approach to automated pulmonary nodule detection from whole-lung low-dose CT scans. In *Medical Imaging 2009: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Lake Buena Vista, FL, USA, 7–12 February 2009; Volume 7260, p. 72601Q.
29. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [[CrossRef](#)]
30. Hall, M.A. *Correlation-Based Feature Selection of Discrete and Numeric Class Machine Learning*. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford, CA, USA, 29 June–2 July 2000; pp. 359–366.
31. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [[CrossRef](#)]
32. Chakroborty, S.; Saha, G. Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification. *Speech Commun.* **2010**, *52*, 693–709. [[CrossRef](#)]
33. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
34. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [[CrossRef](#)]
35. Bersimis, F.G.; Varlamis, I. Use of health-related indices and classification methods in medical data. In *Classification Techniques for Medical Image Analysis and Computer Aided Diagnosis*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 31–66.
36. Sweilam, N.H.; Tharwat, A.; Moniem, N.A. Support vector machine for diagnosis cancer disease: A comparative study. *Egypt. Inform. J.* **2010**, *11*, 81–92. [[CrossRef](#)]
37. Wang, H.; Huang, G. Application of support vector machine in cancer diagnosis. *Med. Oncol.* **2011**, *28*, 613–618. [[CrossRef](#)]
38. Cascio, D.; Taormina, V.; Cipolla, M.; Bruno, S.; Fauci, F.; Raso, G. A multi-process system for HEP-2 cells classification based on SVM. *Pattern Recognit. Lett.* **2016**, *82*, 56–63. [[CrossRef](#)]
39. Cascio, D.; Taormina, V.; Raso, G. Deep convolutional neural network for HEP-2 fluorescence intensity classification. *Appl. Sci.* **2019**, *9*, 408. [[CrossRef](#)]
40. Cascio, D.; Taormina, V.; Raso, G. Deep CNN for IIF images classification in autoimmune diagnostics. *Appl. Sci.* **2019**, *9*, 1618. [[CrossRef](#)]
41. Konukoglu, E.; Glocker, B. Random forests in medical image computing. In *Handbook of Medical Image Computing and Computer Assisted Intervention*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 457–480.
42. Nishio, M.; Nishizawa, M.; Sugiyama, O.; Kojima, R.; Yakami, M.; Kuroda, T.; Togashi, K. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS ONE* **2018**, *13*, e0195875. [[CrossRef](#)]
43. Zhang, X.; Li, T.; Wang, J.; Li, J.; Chen, L.; Liu, C. Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy. *Front. Genet.* **2019**, *10*, 735. [[CrossRef](#)] [[PubMed](#)]
44. Xie, Y.; Meng, W.Y.; Li, R.Z.; Wang, Y.W.; Qian, X.; Chan, C.; Yu, Z.F.; Fan, X.X.; Pan, H.D.; Xie, C.; et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Transl. Oncol.* **2021**, *14*, 100907. [[CrossRef](#)]
45. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
46. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
47. Delzell, D.A.; Magnuson, S.; Peter, T.; Smith, M.; Smith, B.J. Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Front. Oncol.* **2019**, *9*, 1393. [[CrossRef](#)]
48. Parmar, C.; Grossmann, P.; Bussink, J.; Lambin, P.; Aerts, H.J. Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* **2015**, *5*, 1–11. [[CrossRef](#)]
49. Zhang, Y.; Oikonomou, A.; Wong, A.; Haider, M.A.; Khalvati, F. Radiomics-based prognosis analysis for non-small cell lung cancer. *Sci. Rep.* **2017**, *7*, 1–8. [[CrossRef](#)]
50. Sun, T.; Wang, J.; Li, X.; Lv, P.; Liu, F.; Luo, Y.; Gao, Q.; Zhu, H.; Guo, X. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. *Comput. Methods Programs Biomed.* **2013**, *111*, 519–524. [[CrossRef](#)] [[PubMed](#)]
51. El-Baz, A.; Beache, G.M.; Gimel'farb, G.; Suzuki, K.; Okada, K.; Elnakib, A.; Soliman, A.; Abdollahi, B. Computer-aided diagnosis systems for lung cancer: Challenges and methodologies. *Int. J. Biomed. Imaging* **2013**, *2013*, 1–31. [[CrossRef](#)]
52. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [[CrossRef](#)]