Article

Deep Radiotranscriptomics of Non-Small Cell Lung Carcinoma for Assessing Molecular and Histology Subtypes with a Data-Driven Analysis

Eleftherios Trivizakis [1,2,*], John Souglakos [1,3,4], Apostolos Karantanas [1,5] and Kostas Marias [1,6]

1. Computational Biomedicine Laboratory (CBML), Foundation for Research and Technology Hellas (FORTH), 70013 Heraklion, Greece; souglak@uoc.gr (J.S.); karantanas@uoc.gr (A.K.); kmarias@ics.forth.gr (K.M.)
2. Medical School, University of Crete, 71003 Heraklion, Greece
3. Laboratory of Translational Oncology, Medical School, University of Crete, 71003 Heraklion, Greece
4. Department of Medical Oncology, University Hospital of Heraklion, 71500 Heraklion, Greece
5. Department of Radiology, Medical School, University of Crete, 71003 Heraklion, Greece
6. Department of Electrical and Computer Engineering, Hellenic Mediterranean University, 71410 Heraklion, Greece
* Correspondence: trivizakis@ics.forth.gr

Abstract: Radiogenomic and radiotranscriptomic studies have the potential to pave the way for a holistic decision support system built on genomics, transcriptomics, radiomics, deep features and clinical parameters to assess treatment evaluation and care planning. The integration of invasive and routine imaging data into a common feature space has the potential to yield robust models for inferring the drivers of underlying biological mechanisms. In this non-small cell lung carcinoma study, a multi-omics representation comprised deep features and transcriptomics was evaluated to further explore the synergetic and complementary properties of these diverse multi-view data sources by utilizing data-driven machine learning models. The proposed deep radiotranscriptomic analysis is a feature-based fusion that significantly enhances sensitivity by up to 0.174 and AUC by up to 0.22, compared to the baseline single source models, across all experiments on the unseen testing set. Additionally, a radiomics-based fusion was also explored as an alternative methodology yielding radiomic signatures that are comparable to several previous publications in the field of radiogenomics. Furthermore, the machine learning multi-omics analysis based on deep features and transcriptomics achieved an AUC performance of up to $0.831 \pm 0.09/0.925 \pm 0.04$ for the examined molecular and histology subtypes analysis, respectively. The clinical impact of such high-performing models can add prognostic value and lead to optimal treatment assessment by targeting specific oncogenes, namely the response of tyrosine kinase inhibitors of *EGFR* mutated or predicting the chemotherapy resistance of *KRAS* mutated tumors.

Keywords: non-small cell lung carcinoma; radiotranscriptomics; deep features; radiomics; transcriptomics; machine learning; multi-view learning

1. Introduction

The highest mortality rate worldwide has been estimated as being among lung cancer patients, according to a recent report [1] by the World Health Organization (WHO [2]). Therapeutic decisions for non-small cell lung carcinoma (NSCLC) in contemporary clinical practice are based on empirical observations of clinicians in association with histological, genomic, clinical, laboratory and other routine imaging data [3]. Gene expression analysis provides insights into the biological functions and molecular structure of neoplasms, while the profiling of specific malignancies contributes to the discovery of novel and discriminative biomarkers by selecting an optimal and personalized treatment [4]. The molecular characteristics of NSCLC should be considered in treatment decisions as they are involved in the crucial mechanisms of lesion progression [5]. Furthermore, the effectiveness

of radiomics is based on the hypothesis that medical image analysis can quantify the underlying disease. In this context, radiogenomic/radiotranscriptomic analysis [6] has two main goals: (a) the correlation of imaging with genomic/transcriptomic features, and (b) the combination of the aforementioned data sources to improve robustness for increased predictive power. In particular, the accurate prediction of the genetic alterations of targeted oncogenes has a high clinical significance in precision medicine as they have the potential to uncover prognostic drivers for treatment response [7–16].

NSCLC radiogenomic/radiotranscriptomic analyses in the current literature mainly focus on predicting molecular and histological subtypes, solely from imaging data, and correlating genomic signatures with radiomic features [17–26]. Only a handful of studies have combined selected radiomic and transcriptomic features into a unified predictive signature. In particular, radiotranscriptomics of adipose tissue has been used for risk assessment in cardiovascular disease [27]. In oncology, Chaddad et al. [28] performed a variety of multi-omics analyses by integrating radiomics with genomics, transcriptomics, proteomics and clinical data to assess the survival rate of IDH1 wild-type glioblastoma patients. Fan et al. [29] implemented a nomogram-based integration of radiomics, transcriptomics and clinical parameters to estimate the objective response rate, and the overall and progression-free survival of NSCLC patients treated with radiotherapy. Thus, the combination of the two data sources provided a robust and improved model in terms of predictive power [28,29].

In this study, a multi-view analysis was used to combine deep features with selected transcriptomic features in a common multi-omics space to predict the molecular subtypes (*EGFR*, *KRAS* mutation) and histological subtypes (adenocarcinoma or squamous cell carcinoma) of NSCLC patients. Applying domain agnostic and data-driven machine learning techniques to the examined deep radiotranscriptomic data has the advantage of capturing the biological variability of NSCLC and, consequently, improving robustness and prediction performance. Despite the increased popularity of artificial intelligence, to the best of our knowledge, the current work is the first deep radiotranscriptomic analysis of NSCLC that integrates selected transcriptomics and deep features into a unified feature space.

2. Materials and Methods

2.1. Dataset

The NSCLC Radiogenomics [30] dataset comprises 211 CT routine examinations in total, with 142 available ePad [31] pixel-based lesion annotations and 211 image markup standards (AIM files), and an additional 162 PET/CT examinations, 130 RNA-seq vectors (P_G) and clinical data with genomic, histology, semantic, survival or disease recurrence information. The patient cohort of this study includes up to 112 subjects from the examined database, specifically the routine CT scans with the available pixel-based annotations, transcriptomic data and well-defined histology or molecular endpoints. The aforementioned clinical data includes patients with characterizations such as *EGFR* mutation status (L_{EGFR} = 172), *KRAS* mutation status (L_{KRAS} = 171) and histology subtype (L_{HS} = 211). A subset of 142 CT examinations has available annotations on a pixel basis for the region of interest (P_{ROI}). The intersection of the imaging and transcriptomic data, denoted by $P_{ROI} \cap P_T$, is a set of 115 patients (P_{RG}). The final cohorts of *EGFR* ($P_{EGFR} = L_{EGFR} \cap P_{RG}$ = 92), *KRAS* ($P_{KRAS} = L_{KRAS} \cap P_{RG}$ = 93) and histology ($P_{HS} = L_{HS} \cap P_{RG}$ = 112) subtypes were considered for the proposed radiotranscriptomic analyses.

2.2. Multi-View Learning for Radiotranscriptomics

Two data views were considered in this study: (a) deep features or traditional radiomics and (b) transcriptomics. The fusion strategy that was incorporated into the proposed radiotranscriptomic analysis includes the concatenation of both views into a common feature space prior to classification. A depiction of the full radiotranscriptomic pipeline can be found in Figure 1. Details regarding the selected parameters and the complete source code of the analysis are provided online [32].
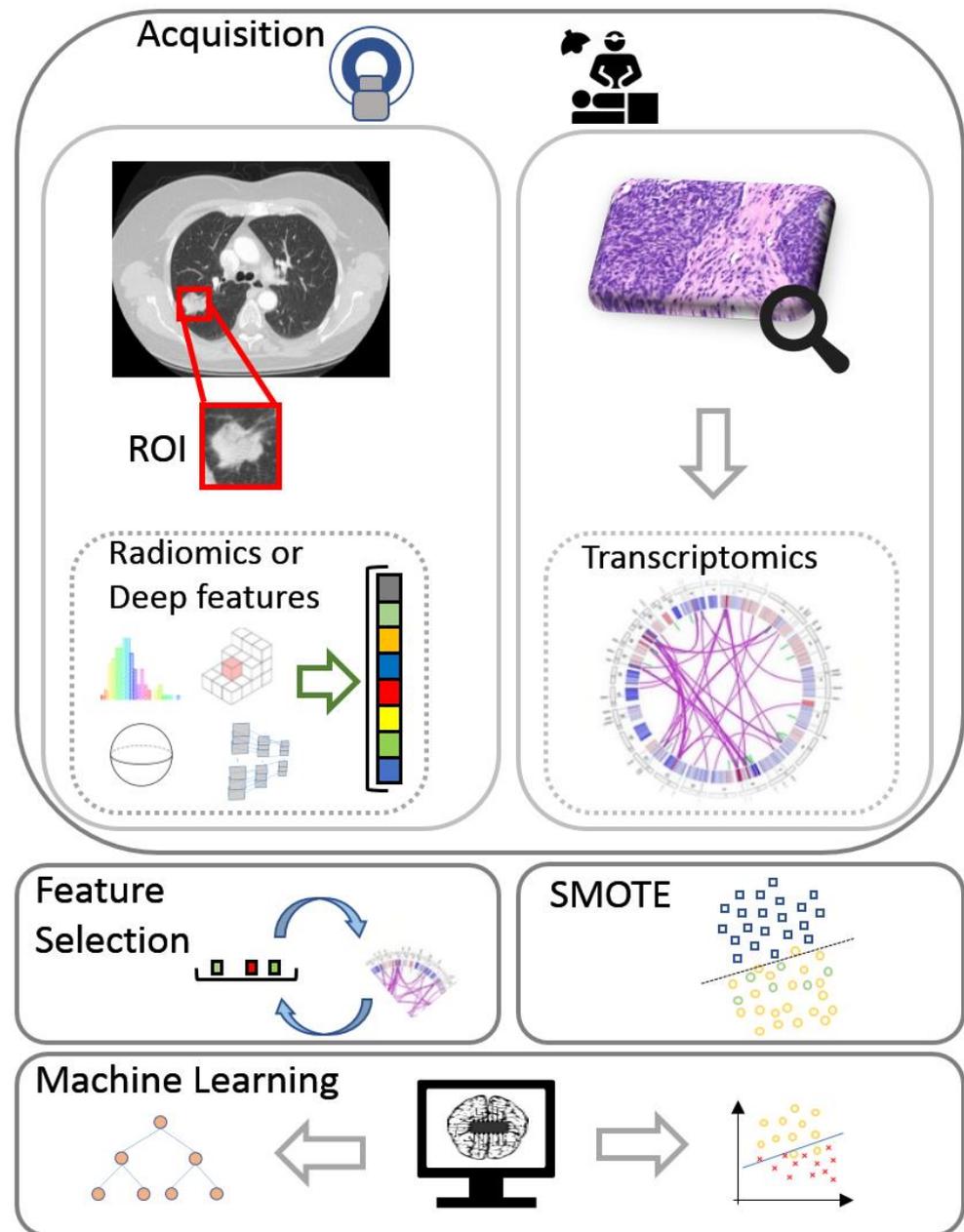
Figure 1. The flow diagram of the proposed radiotranscriptomic analysis incorporates acquisition of transcriptomics and computed tomography data with pixel-wise lesion delineation, followed by feature extraction, feature selection, minority oversampling, multi-view integration, and machine learning analysis. ROI, region of interest; SMOTE, synthetic minority oversampling technique.

2.2.1. Deep Features

Deep learning has a substantial impact on image analysis tasks, primarily because of the deep models' capacity to generalize [33,34]. This is achieved by learning low-level filters that are incorporated into the hierarchical inner representation of the convolutional part of the deep model [35]. Transfer learning (TL) is one of the most critical methodologies in deep learning since it enables the implementation of deeper models without the need for big data availability. Two main types of domain adaptation [36,37] have been proposed: (a) "off-the-shelf" TL, where the feature extraction part of a source model is transferred to the target model; and (b) fine-tuning TL, where the whole or part of the source model is transferred to the new model but the internal representation has to be adapted with a new training process. The latter methodology is more demanding on the dataset samples

for model fitting, validation and evaluation. Thus, the "off-the-shelf" TL was used in the proposed methodology because of the low number of available samples in the examined patient cohort and the unbalanced natural prevalence of the disease. Additionally, this methodology has been successfully integrated into many medical image classification tasks, such as interstitial lung disease [37], colonic polyps [38], breast cancer [39], breast density assessment [40] and brain neoplasms [41], and evaluated across multiple other histopathology datasets [36].

The proposed "off-the-shelf" TL strategy incorporates pretrained ImageNet [42] models extracting raw deep features from the last convolution layer of the source model. Thus, no training was applied to the TL models since the investigated dataset was limited by size, which rendered de novo network development impractical. Eighteen models with a variety of architectures and parameters were used, including the most popular VGG [43], Inception [44], Xception [45], ResNet [46], NasNet [47], MobileNet [48] and DenseNet [49], and their variants that are available in the Keras [50] online repository. All pretrained convolutional layers were transferred to the new model, but the high level fully connected classification layers were removed to allow for the deep feature extraction from the low-level filters. The fully convolutional model was used on a per slice premise and the maximum pooling on a patient basis, resulting in a single compact representation of the three-dimensional volume of interest. Furthermore, the extraction was performed on the zero-padded CT region of interest (ROI) with a size of 150 by 150 pixels. Image normalization was performed prior to the padding. Depending on the architecture used, the number of raw features extracted per slice varied from 1088 to 65,919. Features with zero variance were removed, significantly reducing the length of the extracted vector, and feature standardization was applied for zero mean and unit variance before the analysis.

2.2.2. Radiomics

The radiomic analysis comprised 2996 imaging features extracted with a fixed bin size from the volume of interest of the original CT examination. Shape features (fourteen in total) including elongation, flatness, sphericity, 3D and 2D diameter, mesh, surface and voxel volume were calculated in addition to the first order features (eighteen) of skewness, energy, entropy, kurtosis and other statistical features. Texture features (seventy-five), such as autocorrelation, cluster prominence, contrast, gray-level covariance (GLCM), dependence (GLDM), run length (GLRLM), size zone (GLSZM) and neighborhood gray-tone difference (NGTDM) matrix features, were extracted by the PyRadiomics framework [51] version 2.2.0. Additionally, isotropic resampling was performed using the built-in Pyradiomics preprocessing methods to achieve uniform spacing across patients. Other image filtering techniques (six) were applied to the original examination, including exponential, gradient, Laplacian of Gaussian, square, square root and wavelet filtering (twenty-two) prior to feature extraction, which enriched the proposed radiomic analysis by augmenting the final feature vector. In particular, mother wavelets such as daubechies, symlets, coiflets, biorthogonal and reverse biorthogonal with decomposition up to the second level were applied to the original examination.

2.2.3. Transcriptomics

Transcriptomic data provide details about carcinogenesis procedures and neoplasm progression [52]. Additionally, transcriptomic profiling is a significant technology for improving diagnosis, patient stratification and the identification of prognostic biomarkers [4]. Thus, this personalized transcriptomic evaluation could promote bespoke therapies or response prediction based on the specific neoplasm composition. The examined RNA-seq data were downloaded from the NCBI GEO hosting database [53]. The pseudonyms for the subjects are the same as those used in the image database. In total, 130 RNA-seq vectors were available for radiotranscriptomic analysis. The original transcriptomics comprised 22,126 values but, after removing incomplete features, a transcriptomic signature of 5268 molecules per patient was examined.

2.2.4. Feature Selection

The analysis of variance (ANOVA) was used separately as an objective-specific feature selection process for transcriptomics and deep feature vectors. This supervised univariate method reduced the dimensionality of the extracted feature vectors in a meaningful way according to the examined classification objective during the experimental phase. The resulting statistically significant components were selected for identifying potential markers for differentiating between the mutation status and histology subtypes of the examined patients. A combined analysis of p values with respect to their corresponding F scores for transcriptomic, radiomic and deep features assisted in the selection of a subset of the most significant radiotranscriptomic features. Additionally, an L1 penalty powered by a linear regression model was applied to further reduce the feature space of each data view by minimizing the coefficients, which yielded a sparse representation. The extensive lists of both transcriptomic and radiomic signatures are provided in the Supplementary Materials.

2.2.5. Synthetic Minority Oversampling Technique

A common problem in data analysis is that datasets contain unequal distributions across categories, with "normal" examples outnumbering the uncommon "abnormal" occurrences. This can lead to a negative bias with reduced sensitivity towards the minority class in machine learning classification. The sample generation with SMOTE [54] is a supervised approach that uses the k-nearest neighbors in the feature space to augment the samples with artificial data points. Between two neighbors, a new feature vector is generated and multiplied by a random factor of positive decimal with a value of less than one. SMOTE was applied in the training set to alleviate the imbalances during the model convergence. Additionally, experiments without the SMOTE were performed, maintaining the natural prevalence of the disease. The trained models were evaluated exclusively on "real" and unseen samples.

2.2.6. Data Stratification

Fivefold cross-validation on a patient basis was applied to the original dataset for splitting into training and testing sets. Furthermore, for the examined patient cohort, patient stratification was applied in a way that preserved the balance of each class. The training set was used for model fitting, feature selection and oversampling of the minority class. The class distribution across the corresponding experiments was: (a) 80% wild-type and 20% mutant, (b) 76.1% wild-type and 23.9% mutant and (c) 79.8% adenocarcinoma and 20.2% squamous cell carcinoma. A key factor of the utmost importance during the experimental process was that the testing set remain unseen until the final stage of the performance evaluation, as depicted in Figure 2. This approach was applied to enhance the reliability for all developed machine learning models and avoid the overfitting of the data distribution or sample selection biases.

2.2.7. Classification

Three binary tumor characterizations were examined: (a) EGFR mutation status, (b) KRAS mutation status and (c) histology subtypes. Seven classifiers were employed interchangeably for differentiating among the radiotranscriptomic, transcriptomic and radiomic signatures, namely: (a) k-NN; (b) decision tree; (c) RBF-GPC; (d) RBF-SVM; (e) linear SVM; (f) polynomial SVM; and (g) sigmoid SVM. The classifier implementations of the Sci-Kit Learn library [55] were used in this study.
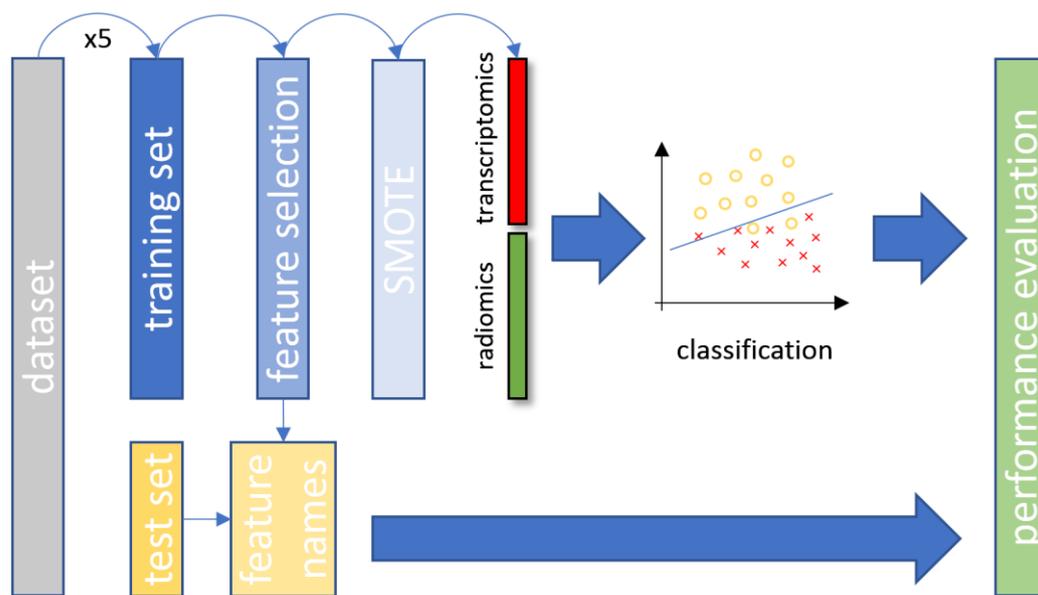
Figure 2. The overall data analysis process with the proposed CT and transcriptomic feature fusion in a combined machine learning analysis. SMOTE, synthetic minority oversampling technique; CT, computed tomography.

3. Results

The analysis was performed on a computational node integrating an AMD Ryzen central processing unit with thirty-two threads, sixty-four gigabytes of random access memory and an RTX 3090 graphics processing unit with twenty-four gigabytes of video memory. Overall, the same experimental protocol and data stratification methodology were applied across all experiments, with the key differentiating factors being the feature fusion (radiomics or deep features with transcriptomics), oversampling technique and classifier. In total, 2394 unique models were evaluated, including deep radiotranscriptomics (756, SMOTE/not by deep models by endpoints by classifier types), traditional radiotranscriptomics (42, SMOTE/not by endpoints by classifier types), single source transcriptomics (798) and imaging models (798).

The comparison of the corresponding radiotranscriptomic against single source models on ROC curves in Figure 3 reveals the performance advantage in favor of the former, with significantly improved robustness and consistency throughout the examined experiments. In particular, improved performance was observed mainly in radiotranscriptomics for molecular and histological subtype differentiation compared to the best corresponding single source models with gains in AUC scores ranging from 0.016 to 0.22 (Tables S1–S4). The proposed radiotranscriptomics methodology achieved the best classification score with an AUC of 0.943 ± 0.03 on the histology subtype characterization with a linear SVM (Table 1). The deep radiotranscriptomics performed better in molecular subtypes, where the combination of the two sources of selected features impacted greatly on the prediction of *KRAS* mutation status (AUC 0.831 ± 0.09), with improved performance and increased prediction stability from the best single source transcriptomic model (AUC 0.611 ± 0.22). It should be noted that the pre-trained ResNet and DenseNet model families provided the best deep features. Additionally, it is worth mentioning that the single source deep features model for assessing *EGFR* and *KRAS* mutation status outperformed their counterparts in traditional radiomic models.
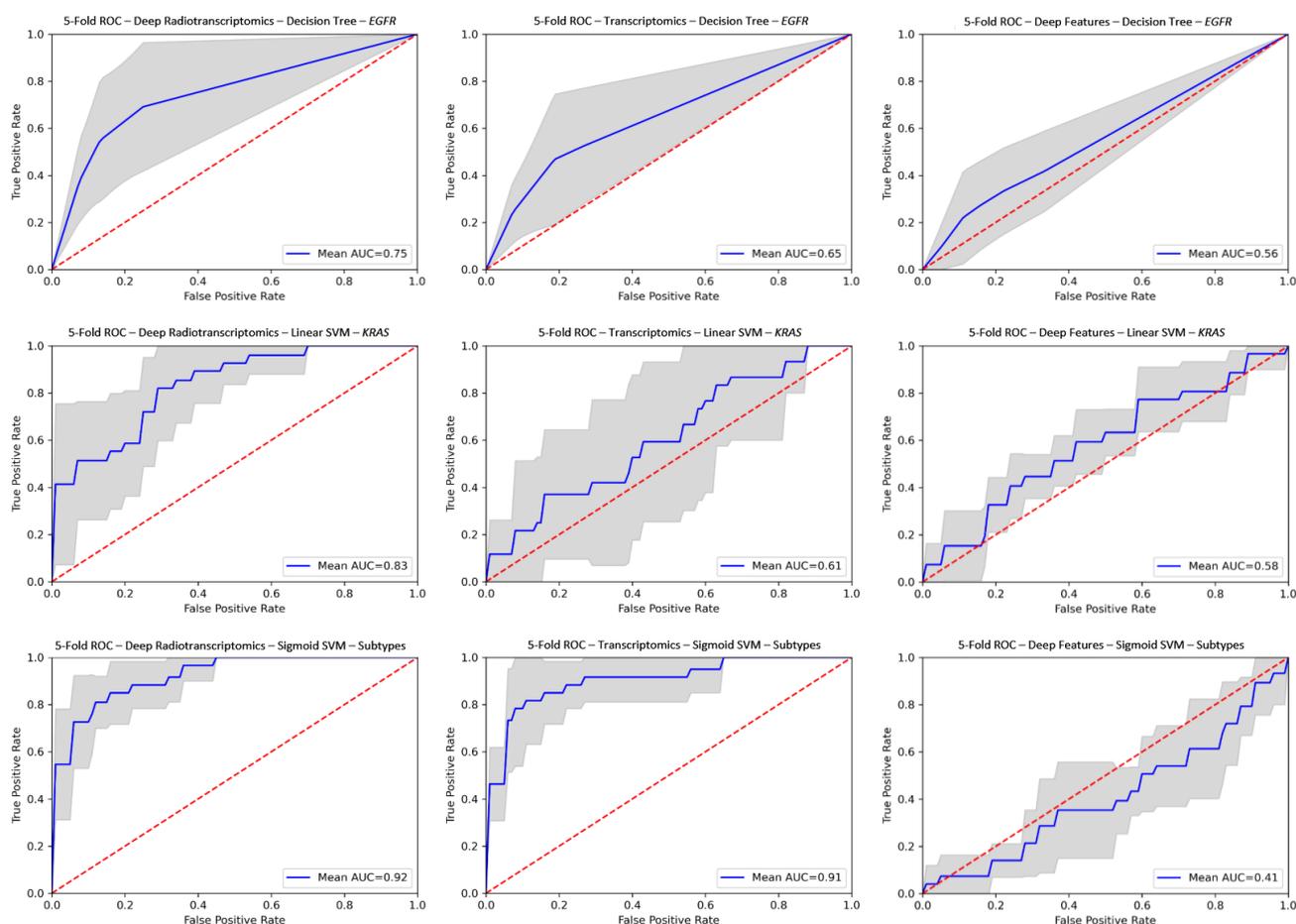
Figure 3. ROC curves for deep radiotranscriptomics (**left column**), transcriptomics (**center column**) and deep feature (**right column**) analysis. The top row represents EGFR (SMOTE), the middle row is KRAS and the bottom row is histology subtypes. The gray region represents the prediction variability among the unseen testing folds. AUC, area under curve; ROC, receiver operating characteristic; EGFR, epidermal growth factor receptor; KRAS, Kristen rat sarcoma; SVM, support vector machine; SMOTE, synthetic minority oversampling technique.

Table 1. Performance of the ML-based pipeline on deep radiotranscriptomics and traditional radiotranscriptomics. The highest overall score between experiments is presented in bold.

Experiments	Classifier	Feature Type	Over-Sampling	ACC	AUC	SN	SPC
EGFR	Decision Tree	ResNet	SMOTE	**0.805 ± 0.05**	**0.747 ± 0.14**	**0.627 ± 0.33**	**0.869 ± 0.06**
KRAS	Linear SVM	DenseNet	No	**0.865 ± 0.08**	**0.831 ± 0.09**	**0.512 ± 0.25**	**0.974 ± 0.03**
Histology Subtypes	Sigmoid SVM	ResNet	No	0.888 ± 0.07	0.925 ± 0.04	0.743 ± 0.16	0.933 ± 0.06
EGFR	Sigmoid SVM	Radiomics-based	SMOTE	0.761 ± 0.10	0.726 ± 0.10	0.600 ± 0.16	0.800 ± 0.11
KRAS	Linear SVM		No	0.730 ± 0.05	0.719 ± 0.07	0.34 ± 0.27	0.883 ± 0.08
Histology Subtypes	Linear SVM		No	**0.907 ± 0.05**	**0.943 ± 0.03**	**0.797 ± 0.12**	**0.941 ± 0.03**

The inclusion of SMOTE in the analysis considerably improved both deep and traditional radiotranscriptomics for *EGFR* expression prediction, especially in terms of sensitivity. In particular, the traditional radiotranscriptomics achieved an AUC of 0.726 ± 0.10 (Table S2) over 0.645 ± 0.12 (Table S1) and the deep feature-based counterpart achieved an AUC of 0.747 ± 0.14 (Table S4) over 0.634 ± 0.24 (Table S3).

The traditional radiotranscriptomics for *EGFR* performed slightly better (AUC 0.645 ± 0.12 versus 0.642 ± 0.11) than the best single source counterpart, according to Table S1. This difference was negligible for the aforementioned model, with the prediction AUC only

increasing by 0.003 compared to the transcriptomics model. Although, in terms of sensitivity and specificity, the difference is more pronounced with a difference of 0.04 and 0.043, respectively.

The most discriminative components in the radiomic signature for *EGFR* include features that reflect similarity, homogeneity, heterogeneity and complexity in texture patterns (gldm_DependenceNonUniformityNormalized, glszm_ZoneEntropy, glcm_MCC (maximal correlation coefficient), ngtdm_Strength). The *KRAS* radiomic signature comprises features that quantify skewness, asymmetry, local homogeneity and substantial variations in intensity values in the region of interest (glcm_ClusterShade, glcm_Idmn (inverse difference moment normalized), gldm_LargeDependenceLowGrayLevelEmphasis, ngtdm_Complexity). The radiomics of histology subtypes are mainly zone-based features that estimate the variability of zone size, the ratio of large and small zones to high gray-levels and skewness in texture complexity (glszm_SizeZoneNonUniformityNormalized, glszm_ZonePercentage, glszm_LargeAreaHighGrayLevelEmphasis, glszm_SmallAreaHighGrayLevelEmphasis, glcm_MCC), along with a few first order features such as the total energy, 90th percentile and wavelet minimum. It is important to note that at least half of the features in each radiomic signature were based on the wavelet filtered image. Furthermore, two detailed lists of the most discriminative features for both transcriptomics and radiomics in molecular or histological subtypes are shown in Tables S5 and S6. Detailed performance metrics of all the experiments are presented in Tables 1 and S1–S4 and the corresponding ROC curves in Figure 3 and Tables S2–S6.

4. Discussion

Artificial intelligence has advanced into an essential methodology for inferring knowledge from a high dimensional space with a data-driven perspective in many disciplines. In medicine, the increasing quantities of information could outline the complexity of the underlining biology of specific lesions, especially in oncology. While several efforts have used single source data to investigate and model cancer mechanisms [56–60], our effort is towards the synergistic use of high dimensional and high throughput data (deep features, radiomics and transcriptomics) for identifying the prognostic signatures towards precision decision support in oncology.

4.1. Common Features Found in Current Literature

The proposed approach yielded distinct signatures for *EGFR*, *KRAS* and histology subtypes across all three experiment types, as illustrated in Tables S5 and S6. Notably, the most significant radiomic feature in this analysis for *KRAS* mutation was the ngtdm_Complexity, which is the same feature found by Zhang et al. [61] for the same radiogenomic objective in a completely different patient cohort, indicating feature stability and robustness. Additionally, in the same study, the best radiomic feature for *EGFR* mutation assessment was the gldm_LargeDependenceHighGrayLevelEmphasis, which belongs to the same texture family as gldm_DependenceNonUniformityNormalized and is similar in nature to the high level feature glrlm_LongRunHighGrayLevelEmphasis, both also found in the present analysis. Furthermore, another agreement with the findings (Table S6) of the proposed analysis was observed for discriminative features of histology subtypes with studies reporting cluster shade [62], first-order, GLCM, GLSZM [63] and a combination of high level emphasis and small area emphasis [64] as important features. It is worth noting that the majority of the identified radiomics were wavelet-based features (Table S6), indicating that a significant part of the differentiating information exists only in specific frequency bands and can be deciphered through scale-space wavelet analysis [65].

4.2. Performance of Radiotranscriptomics Versus Single Source Models

The single source radiomics and deep features analyses achieved a performance at the lower end of the spectrum in terms of sensitivity, but both have shown increased specificity compared to the corresponding transcriptomic analyses. The deep features achieved

a higher sensitivity in *EGFR/KRAS* mutation status assessment than the corresponding radiomics. The prediction of the molecular and histology subtypes was enhanced by the integration of imaging with transcriptomic data in a common feature space. The proposed radiotranscriptomics established models with an improved sensitivity of up to 0.182 compared to their single source counterparts, a superior AUC performance in mutation status prediction of up to 0.831 ± 0.09 and a histological subtype prediction of up to 0.942 ± 0.03. These results showcase the radiotranscriptomic synergy assumption between the two different sources of data, as discussed in the introduction. Furthermore, the benefits of radiotranscriptomic analysis can be summarized into the overall improved ROC curves (Figure 3), AUC score and sensitivity, and the reduced prediction variability in classification across all types of experiments (*EGFR*, *KRAS* and histology subtypes). In comparison, similar studies on imaging only data demonstrate significantly lower performance for *KRAS* mutation prediction, up to AUC 0.667 [23], and histological subtype differentiation ranging from 0.754 [21] to 0.893 [26]. In contrast, the radiotranscriptomics of *EGFR* did not improve on the current state-of-the-art study of Rizzo et al. [23], which was based solely on semantic CT features. This can also be attributed to the imaging modality, since better and more discriminative features or biomarkers [66] have been reported in several studies [25,67–69] where the PET/CT radiomic signature with an AUC of 0.805 significantly outperformed the CT only features (AUC 0.667) in *EGFR* mutation status differentiation. Despite the limited number of patients for machine learning analysis, especially regarding the molecular subtype patient cohorts, the proposed radiotranscriptomic model for *KRAS* differentiation outperformed the model of Rizzo et al. [23]. Additionally, the proposed ML-based analysis for molecular and histological subtypes outperformed the corresponding NSCLC state-of-the-art research [17,19–23,25,26] by a wide margin. The performance of the proposed radiotranscriptomics and the state-of-the-art literature is presented in detail in Table 2. The complete results of the radiotranscriptomic analyses are presented in Table 1 and Tables S1–S4 of the Supplementary Materials.

Table 2. The corresponding literature of the examined dataset with varying methodologies including semantic CT features, radiomic and radiotranscriptomics analyses (AUC). The highest overall score for each experiment type is presented in bold.

	EGFR	*KRAS*	Histological Subtypes
Proposed Traditional Radiotranscriptomics	0.726 ± 0.10	0.719 ± 0.07	**0.942 ± 0.03**
Proposed Deep Radiotranscriptomics	**0.747 ± 0.14**	0.831 ± 0.09	0.924 ± 0.04
Morgado et al. [17]	0.737	-	-
Moreno et al. [19]	up to 0.82	up to 0.778	-
Dong et al. [20]	0.751	0.696	-
Yamada et al. [21]	-	-	0.754
Koyasu et al. [22]	0.659	-	0.843
Rizzo et al. [23]	0.823	0.667	-
Li et al. [25]	0.667	-	-
Zhu et al. [26]	-	-	0.893

4.3. Clinical Impact of the Study

The proposed analyses in this study identified discriminative compact transcriptomics, radiomics and deep feature signatures to accurately model the underlying biology of non-small cell lung cancer. A multi-view learning methodology for high dimensional and low sample sized datasets [70] is essential for the integration of the different types of omics data. Two views that include deep features and transcriptomics were analyzed using distinct pipelines and were integrated prior to classification, with the purpose of capturing the heterogeneity of NSCLC. The high performance of the proposed deep radiotranscriptomics

in assessing the genetic alterations of *EGFR* and *KRAS* oncogenes, as presented in the results section, could potentially add precision in treatment planning with TKIs or other targeted therapies.

Some challenges for single source RNA-seq data that significantly affect the generalization of the analysis are the varying data acquisition protocols [71], intratumor heterogeneity [72] and local mutation burden [73]. These are prominent aspects of non-small cell lung carcinomas. Transcriptomics is subjected to local mutational diversity, which can be as high as interpersonal variation [73]. Imaging features, on the other hand, are computed throughout the whole tumor, yielding complex signatures that include patterns from the microenvironment (necrotic, hypoxic and oxygenated tissue) of the neoplasm. Radiotranscriptomics has the potential to accurately capture the total mutational burden (TMB) by combining locally (transcriptomics) and globally (radiomics) dependent patterns. Therefore, the proposed composite signature addresses the shortcoming of single source data by assembling a holistic representation that fuses markers related to biological mechanisms (via transcriptomics) and tumor heterogeneity patterns (via imaging features extracted from the volume of interest). This can be beneficial for assessing clinically impactful endpoints, such as survival and therapy response [73]. To this end, we provide an open source repository [32] of the proposed methodology to encourage further research and to enable reproducibility and comparison with future studies.

4.4. Limitations and Future Extensions

A significant limitation of this study was the reduced data variability from an ethnically diverse patient cohort that reflects the genetic heterogeneity in non-small cell lung carcinomas. This is a key factor for enhancing the predictive power and robustness of radiotranscriptomic models. Another issue was the limited simultaneous availability of both imaging and transcriptomic data. The small dataset size increased prediction variability in many experiments, particularly in some transcriptomics and radiomics for molecular subtype models. The standardization, robustness and reproducibility of radiomics is an important issue in AI analysis that requires multiple examinations of the studied patient cohort. Nonetheless, CT scans use Hounsfield Units to capture the tissue's electron density inside a particular voxel, which is a constant value for each tissue type, offering a unique radiation absorption signature and making CT quantitative by nature. The variability in the genetic and phenotypic profile of NSCLC tumors, which makes radiomic and transcriptomic analysis difficult, was another limiting factor. This can be observed in Table 1 where the standard deviation of the experiments can reach up to 0.22 in terms of AUC variability. As a result, parameter optimization in the machine learning methods that was used was especially challenging since it was difficult to assess the impact of the tuning. Therefore, the default parameters were used for every component of the pipeline, despite the feature type. Additionally, instead of a unified feature space, fusion could be achieved at the decision level with different classifiers or analysis pipelines for each data type. Finally, another limiting factor was the lack of the multi-modal fusion of imaging features because, while PET/CT examinations are available in this dataset, calculating radiomic and deep features was not feasible due to the lack of pixel-based annotations. The latter requires at least two expert clinicians to perform the delineation and achieve a consensus in terms of tumor margins.

Lastly, future research with a multi-institutional and independent patient cohort is a necessary step to evaluate the current experimental procedure by offering increased variety in the examined data sources through integrating different imaging equipment and genome extraction protocols. Additionally, an extension of the proposed data analysis methodology will be explored on the same patient cohort, including endpoints that rely on statistically depended clinical variables, such as lesion recurrence, overall survival or novel therapy-related markers, in order to adapt the therapeutic strategy [74,75] according to the patient's radiogenomic profile into a personalized diagnosis and treatment plan. Other important molecular indicators for assessing targeted treatments include *ALK* rearrange-

ments, *BRAF* mutation status and programmed death-ligand 1 (*PDL1*) expression, which will be investigated in a future radiogenomic study.

5. Conclusions

The deep radiotranscriptomics framework achieved state-of-the-art performance and, most importantly, improved classification metrics, such as AUC and sensitivity, compared to the baseline single source models in the examined molecular and histology subtype analyses. The proposed machine learning selection and feature fusion provided significant evidence supporting the hypothesis presented in the introduction regarding the complementary nature of the two feature types in the context of radiotranscriptomics. A closer collaboration between physicians and data scientists is essential for developing a trustworthy and explainable AI framework that aims to minimize erroneous diagnosis and optimize the planning of a personalized treatment strategy.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/diagnostics11122383/s1: Figure S1, Spearman's correlations (rho > 0.6) among radiomic and transcriptomic features. No statistically significant radiogenomic correlations were found; Figure S2, ROC curves for radiotranscriptomics- (left column), transcriptomics- (center column) and radiomics-based (right column) analysis. The top row represents EGFR (SMOTE), the middle row is KRAS and the bottom row is histology subtypes. The gray region represents the prediction variability among the unseen testing folds. AUC, area under curve, ROC, receiver operating characteristic, EGFR, epidermal growth factor receptor, KRAS, Kristen rat sarcoma, SVM, support vector machine, SMOTE, synthetic minority oversampling technique; Figure S3, ROC curves for radiotranscriptomics- (left column), transcriptomics- (center column) and radiomics-based (right column) analysis. The top row represents EGFR, the middle row is KRAS and the bottom row is histology subtypes; Figure S4, ROC curves for SMOTE radiotranscriptomics- (left column), transcriptomics- (center column) and radiomics-based (right column) analysis. The top row represents EGFR, the middle row is KRAS and the bottom row is histology subtypes; Figure S5, ROC curves for deep radiotranscriptomics- (left column), transcriptomics- (center column) and deep descriptor-based (right column) analysis. The top row represents EGFR, the middle row is KRAS and the bottom row is histology subtypes; Figure S6, ROC curves for SMOTE deep radiotranscriptomics- (left column), transcriptomics- (center column) and deep descriptor-based (right column) analysis. The top row represents EGFR, the middle row is KRAS and the bottom row is histology subtypes; Table S1, Performance of the ML-based pipeline on CT 3D radiomics, transcriptomics and radiotranscriptomics analysis. The highest overall score is presented in bold; Table S2, Performance of the SMOTE ML-based pipeline on CT 3D radiomics, transcriptomics and radiotranscriptomics analysis. The highest overall score is presented in bold; Table S3, Performance of the ML-based pipeline on deep descriptors, transcriptomics and deep radiotranscriptomics analysis. The highest overall score is presented in bold; Table S4, Performance of the SMOTE ML-based pipeline on deep descriptors, transcriptomics and deep radiotranscriptomics analysis. The highest overall score is presented in bold; Table S5, The most significant transcriptomic features; Table S6, The most significant radiomic features.

Author Contributions: E.T. conceived and designed the study. E.T. and K.M. contributed to the performed analysis and interpretation of data and drafted the manuscript. E.T., J.S., A.K. and K.M. contributed to the literature research, interpretation of data and revised the manuscript. J.S. and A.K. contributed to the clinical aspects as well in the critical revision of the paper. K.M. contributed to the critical revision of the paper and was the guarantor of the integrity of the entire study. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The examined computed tomography and transcriptomic dataset titled "NSCLC Radiogenomics" is available online as an open-access repository via the following

link: https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics (accessed on 31 March 2019).

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

References

1. International Agency for Research on Cancer. *Latest Global Cancer Data: Cancer Burden Rises to 18.1 Million New Cases and 9.6 Million Cancer Deaths in 2018*; IARC: Geneva, Switzerland, 2018.
2. World Health Organization. Available online: https://covid19.who.int/ (accessed on 1 August 2020).
3. Potti, A.; Mukherjee, S.; Petersen, R.; Dressman, H.K.; Bild, A.; Koontz, J.; Kratzke, R.; Watson, M.A.; Kelley, M.; Ginsburg, G.S.; et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.* **2006**, *355*, 570–580. [CrossRef]
4. Sanchez-Palencia, A.; Gomez-Morales, M.; Gomez-Capilla, J.A.; Pedraza, V.; Boyero, L.; Rosell, R.; Fárez-Vidal, M.E. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer* **2011**, *129*, 355–364. [CrossRef] [PubMed]
5. Li, J.; Bi, L.; Shi, Z.; Sun, Y.; Lin, Y.; Shao, H.; Zhu, Z. RNA-Seq analysis of non-small cell lung cancer in female never-smokers reveals candidate cancer-associated long non-coding RNAs. *Pathol. Res. Pract.* **2016**, *212*, 549–554. [CrossRef] [PubMed]
6. Trivizakis, E.; Papadakis, G.Z.; Souglakos, I.; Papanikolaou, N.; Koumakis, L.; Spandidos, D.A.; Tsatsakis, A.; Karantanas, A.H.; Marias, K. Artificial intelligence radiogenomics for advancing precision and effectiveness in oncologic care (Review). *Int. J. Oncol.* **2020**, *57*, 43–53. [CrossRef] [PubMed]
7. Yoon, H.-Y.; Ryu, J.-S.; Sim, Y.S.; Kim, D.; Lee, S.Y.; Choi, J.; Park, S.; Ryu, Y.J.; Lee, J.H.; Chang, J.H. Clinical significance of EGFR mutation types in lung adenocarcinoma: A multi-centre Korean study. *PLoS ONE* **2020**, *15*, e0228925. [CrossRef]
8. Aguilar, K.M.; Winfree, K.B.; Muehlenbein, C.E.; Zhu, Y.E.; Wilson, T.; Wetmore, S.; Nadler, E.S. Treatment Patterns by EGFR Mutation Status in Non-Small Cell Lung Cancer Patients in the USA: A Retrospective Database Analysis. *Adv. Ther.* **2018**, *35*, 1905–1919. [CrossRef]
9. Hammoudeh, Z.A.; Antonova, O.; Staneva, R.; Nikolova, D.; Kyuchukov, Y.; Penev, A.; Mintchev, T.; Koleva, V.; Hadjidekova, S.; Toncheva, D. Detecting EGFR mutations in patients with non-small cell lung cancer. *Balk. J. Med. Genet.* **2018**, *21*, 13–18. [CrossRef]
10. Bironzo, P.; Reale, M.L.; Sperone, T.; Tabbò, F.; Caglio, A.; Listì, A.; Passiglia, F.; Di Maio, M.; Righi, L.; Bussolino, F.; et al. Clinical and molecular features of epidermal growth factor receptor (Egfr) mutation positive non-small-cell lung cancer (nsclc) patients treated with tyrosine kinase inhibitors (tkis): Predictive and prognostic role of co-mutations. *Cancers* **2021**, *13*, 2425. [CrossRef]
11. Greenhalgh, J.; Dwan, K.; Boland, A.; Bates, V.; Vecchio, F.; Dundar, Y.; Jain, P.; Green, J.A. First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. *Cochrane Database Syst. Rev.* **2016**. [CrossRef] [PubMed]
12. Borghaei, H.; Paz-Ares, L.; Horn, L.; Spigel, D.R.; Steins, M.; Ready, N.E.; Chow, L.Q.; Vokes, E.E.; Felip, E.; Holgado, E.; et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N. Engl. J. Med.* **2015**, *373*, 1627–1639. [CrossRef] [PubMed]
13. Suda, K.; Tomizawa, K.; Mitsudomi, T. Biological and clinical significance of KRAS mutations in lung cancer: An oncogenic driver that contrasts with EGFR mutation. *Cancer Metastasis Rev.* **2010**, *29*, 49–60. [CrossRef] [PubMed]
14. Román, M.; Baraibar, I.; López, I.; Nadal, E.; Rolfo, C.; Vicent, S.; Gil-Bazo, I. KRAS oncogene in non-small cell lung cancer: Clinical perspectives on the treatment of an old target. *Mol. Cancer* **2018**, *17*, 33. [CrossRef]
15. Massarelli, E.; Varella-Garcia, M.; Tang, X.; Xavier, A.C.; Ozburn, N.C.; Liu, D.D.; Bekele, B.N.; Herbst, R.S.; Wistuba, I.I. KRAS mutation is an important predictor of resistance to therapy with epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. *Clin. Cancer Res.* **2007**, *13*, 2890–2896. [CrossRef]
16. Yang, H.; Liang, S.Q.; Schmid, R.A.; Peng, R.W. New horizons in KRAS-mutant lung cancer: Dawn after darkness. *Front. Oncol.* **2019**, *9*, 953. [CrossRef] [PubMed]
17. 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. [CrossRef]
18. Zhu, Y.; Guo, Y.-B.; Xu, D.; Zhang, J.; Liu, Z.-G.; Wu, X.; Yang, X.-Y.; Chang, D.-D.; Xu, M.; Yan, J.; et al. A computed tomography (CT)-derived radiomics approach for predicting primary co-mutations involving TP53 and epidermal growth factor receptor (EGFR) in patients with advanced lung adenocarcinomas (LUAD). *Ann. Transl. Med.* **2021**, *9*, 545. [CrossRef]
19. Moreno, S.; Bonfante, M.; Zurek, E.; Cherezov, D.; Goldgof, D.; Hall, L.; Schabath, M. A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC. *Tomography* **2021**, *7*, 154. [CrossRef]
20. Dong, Y.; Hou, L.; Yang, W.; Han, J.; Wang, J.; Qiang, Y.; Zhao, J.; Hou, J.; Song, K.; Ma, Y.; et al. Multi-channel multi-task deep learning for predicting EGFR and KRAS mutations of non-small cell lung cancer on CT images. *Quant. Imaging Med. Surg.* **2021**, *11*, 2354–2375. [CrossRef] [PubMed]
21. Yamada, M.; Arimura, H.; Ninomiya, K.; Soufi, M. Automated classification of histological subtypes of NSCLC using support vector machines with radiomic features. In Proceedings of the International Forum on Medical Imaging in Asia 2019; Fujita, H., Lin, F., Kim, J.H., Eds.; SPIE: Bellingham, WA, USA, 2019; Volume 11050, p. 37.
22. Koyasu, S.; Nishio, M.; Isoda, H.; Nakamoto, Y.; Togashi, K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. *Ann. Nucl. Med.* **2020**, *34*, 49–57. [CrossRef] [PubMed]

23. Rizzo, S.; Petrella, F.; Buscarino, V.; De Maria, F.; Raimondi, S.; Barberis, M.; Fumagalli, C.; Spitaleri, G.; Rampinelli, C.; De Marinis, F.; et al. CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer. *Eur. Radiol.* **2016**, *26*, 32–42. [CrossRef] [PubMed]

24. Wang, H.; Schabath, M.B.; Liu, Y.; Stringfield, O.; Balagurunathan, Y.; Heine, J.J.; Eschrich, S.A.; Ye, Z.; Gillies, R.J. Association Between Computed Tomographic Features and Kirsten Rat Sarcoma Viral Oncogene Mutations in Patients with Stage I Lung Adenocarcinoma and Their Prognostic Value. *Clin. Lung Cancer* **2016**, *17*, 271–278. [CrossRef] [PubMed]

25. Li, X.; Yin, G.; Zhang, Y.; Dai, D.; Liu, J.; Chen, P.; Zhu, L.; Ma, W.; Xu, W. Predictive Power of a Radiomic Signature Based on 18F-FDG PET/CT Images for EGFR Mutational Status in NSCLC. *Front. Oncol.* **2019**, *9*, 1062. [CrossRef] [PubMed]

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

27. Oikonomou, E.K.; Williams, M.C.; Kotanidis, C.P.; Desai, M.Y.; Marwan, M.; Antonopoulos, A.S.; Thomas, K.E.; Thomas, S.; Akoumianakis, I.; Fan, L.M.; et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. *Eur. Heart J.* **2019**, *40*, 3529–3543. [CrossRef] [PubMed]

28. Chaddad, A.; Daniel, P.; Sabri, S.; Desrosiers, C.; Abdulkarim, B. Integration of Radiomic and Multi-omic Analyses Predicts Survival of Newly Diagnosed IDH1 Wild-Type Glioblastoma. *Cancers* **2019**, *11*, 1148. [CrossRef]

29. Fan, L.; Cao, Q.; Ding, X.; Gao, D.; Yang, Q.; Li, B. Radiotranscriptomics signature-based predictive nomograms for radiotherapy response in patients with nonsmall cell lung cancer: Combination and association of CT features and serum miRNAs levels. *Cancer Med.* **2020**, *9*, 5065–5074. [CrossRef]

30. Bakr, S.; Gevaert, O.; Echegaray, S.; Ayers, K.; Zhou, M.; Shafiq, M.; Zheng, H.; Benson, J.A.; Zhang, W.; Leung, A.N.C.; et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **2018**, *5*, 180202. [CrossRef]

31. Rubin, D.L.; Akdogan, M.U.; Altindag, C.; Alkim, E. ePAD: An Image Annotation and Analysis Platform for Quantitative Imaging. *Tomography* **2019**, *5*, 170–183. [CrossRef] [PubMed]

32. Trivizakis, E. NSCLC Radiogenomics Source Code. Available online: https://github.com/trivizakis/nsclc_radiotranscriptomics (accessed on 8 June 2021).

33. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; Lehmann, P.H., Ed.; IEEE: Piscataway, NJ, USA, 2014; Volume 7389, pp. 512–519.

34. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 1717–1724. [CrossRef]

35. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.

36. Mormont, R.; Geurts, P.; Maree, R. Comparison of Deep Transfer Learning Strategies for Digital Pathology. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2343–234309.

37. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]

38. Ribeiro, E.; Uhl, A.; Wimmer, G.; Häfner, M. Transfer Learning for Colonic Polyp Classification Using Off-the-Shelf CNN Features. In *Computer-Assisted and Robotic Endoscopy*; Springer: Cham, Switzerland, 2016; Volume 10170, pp. 1–13. [CrossRef]

39. Zhi, W.; Wing, H.; Yueng, F.; Chen, Z.; Zandavi, S.M.; Lu, Z.; Chung, Y.Y. Using Transfer Learning with Convolutional Neural Networks to Diagnose Breast Cancer from Histopathological Images. In Proceedings of the 24th International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017. [CrossRef]

40. Giannakakis, G.; Trivizakis, E.; Tsiknakis, M.; Marias, K. A novel multi-kernel 1D convolutional neural network for stress recognition from ECG. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, UK, 3–6 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.

41. Ioannidis, G.S.; Trivizakis, E.; Metzakis, I.; Papagiannakis, S.; Lagoudaki, E.; Marias, K. Pathomics and Deep Learning Classification of a Heterogeneous Fluorescence Histology Image Dataset. *Appl. Sci.* **2021**, *11*, 3796. [CrossRef]

42. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.

43. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.

44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 2818–2826.

45. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2017; pp. 1800–1807.

46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 770–778.

47. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning Transferable Architectures for Scalable Image Recognition. *arXiv* **2017**, arXiv:1707.07012.

48. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

49. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993.

50. Chollet, F. Others Keras. Available online: https://keras.io (accessed on 8 June 2021).

51. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J.W.L. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef] [PubMed]

52. Zhang, J.; Zhu, N.; Chen, X. A novel long noncoding RNA LINC01133 is upregulated in lung squamous cell cancer and predicts survival. *Tumor Biol.* **2015**, *36*, 7465–7471. [CrossRef]

53. Bakr, S.; Gevaert, O.; Plevritis, S.K. Identification of Relationships between Molecular and Imaging Phenotypes in Non-Small Cell Lung Cancer Using Radiogenomics Map. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103584 (accessed on 1 January 2021).

54. 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]

55. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

56. Mielgo-Rubio, X.; Calvo, V.; Luna, J.; Remon, J.; Martín, M.; Berraondo, P.; Jarabo, J.R.; Higuera, O.; Conde, E.; De Castro, J.; et al. Immunotherapy Moves to the Early-Stage Setting in Non-Small Cell Lung Cancer: Emerging Evidence and the Role of Biomarkers. *Cancers* **2020**, *12*, 3459. [CrossRef]

57. Ntzifa, A.; Kotsakis, A.; Georgoulias, V.; Lianidou, E. Detection of EGFR Mutations in Plasma cfDNA and Paired CTCs of NSCLC Patients before and after Osimertinib Therapy Using Crystal Digital PCR. *Cancers* **2021**, *13*, 2736. [CrossRef]

58. Alhaddad, L.; Pustovalova, M.; Blokhina, T.; Chuprov-Netochin, R.; Osipov, A.N.; Leonov, S. IR-Surviving NSCLC Cells Exhibit Different Patterns of Molecular and Cellular Reactions Relating to the Multifraction Irradiation Regimen and p53-Family Proteins Expression. *Cancers* **2021**, *13*, 2669. [CrossRef]

59. Hu, W.; Wang, G.; Yarmus, L.B.; Wan, Y. Combined Methylome and Transcriptome Analyses Reveals Potential Therapeutic Targets for EGFR Wild Type Lung Cancers with Low PD-L1 Expression. *Cancers* **2020**, *12*, 2496. [CrossRef] [PubMed]

60. Soundararajan, R.; Fradette, J.J.; Konen, J.M.; Moulder, S.; Zhang, X.; Gibbons, D.L.; Varadarajan, N.; Wistuba, I.I.; Tripathy, D.; Bernatchez, C.; et al. Targeting the Interplay between Epithelial-to-Mesenchymal-Transition and the Immune System for Effective Immunotherapy. *Cancers* **2019**, *11*, 714. [CrossRef] [PubMed]

61. Zhang, T.; Xu, Z.; Liu, G.; Jiang, B.; de Bock, G.H.; Groen, H.J.M.; Vliegenthart, R.; Xie, X. Simultaneous identification of egfr, kras, erbb2, and tp53 mutations in patients with non-small cell lung cancer by machine learning-derived three-dimensional radiomics. *Cancers* **2021**, *13*, 1814. [CrossRef]

62. Yang, F.; Chen, W.; Wei, H.; Zhang, X.; Yuan, S.; Qiao, X.; Chen, Y.W. Machine Learning for Histologic Subtype Classification of Non-Small Cell Lung Cancer: A Retrospective Multicenter Radiomics Study. *Front. Oncol.* **2021**, *10*, 608598. [CrossRef] [PubMed]

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

64. Liu, S.; Liu, S.; Zhang, C.; Yu, H.; Liu, X.; Hu, Y.; Xu, W.; Tang, X.; Fu, Q. Exploratory Study of a CT Radiomics Model for the Classification of Small Cell Lung Cancer and Non-small-Cell Lung Cancer. *Front. Oncol.* **2020**, *10*, 1268. [CrossRef] [PubMed]

65. Trivizakis, E.; Ioannidis, G.S.; Souglakos, I.; Karantanas, A.H.; Tzardi, M.; Marias, K. A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis. *Sci. Rep.* **2021**, *11*, 15546. [CrossRef]

66. Bianconi, F.; Palumbo, I.; Spanu, A.; Nuvoli, S.; Fravolini, M.L.; Palumbo, B. PET/CT radiomics in lung cancer: An overview. *Appl. Sci.* **2020**, *10*, 1718. [CrossRef]

67. Wang, Y.; Han, R.; Wang, Q.; Zheng, J.; Lin, C.; Lu, C.; Li, L.; Chen, H.; Jin, R.; He, Y. Biological significance of18F-FDG PET/CT maximum standard uptake value for predicting egfr mutation status in non-small cell lung cancer patients. *Int. J. Gen. Med.* **2021**, *14*, 347–356. [CrossRef] [PubMed]

68. Sun, X.; Xiao, Z.; Chen, G.; Han, Z.; Liu, Y.; Zhang, C.; Sun, Y.; Song, Y.; Wang, K.; Fang, F.; et al. A PET imaging approach for determining EGFR mutation status for improved lung cancer patient management. *Sci. Transl. Med.* **2018**, 10. [CrossRef] [PubMed]

69. 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 CT Image Using Deep Learning. *Eur. Respir. J.* **2019**, *53*, 1800986. [CrossRef] [PubMed]

70. Cao, H.; Bernard, S.; Sabourin, R.; Heutte, L. Random forest dissimilarity based multi-view learning for Radiomics application. *Pattern Recognit.* **2019**, *88*, 185–197. [CrossRef]

71. Krzak, M.; Raykov, Y.; Boukouvalas, A.; Cutillo, L.; Angelini, C. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. *Front. Genet.* **2019**, *10*, 1253. [CrossRef]

72. Cui, W.; Xue, H.; Wei, L.; Jin, J.; Tian, X.; Wang, Q. High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis. *Hum. Genom.* **2021**, *15*, 7. [CrossRef]

73. Jia, Q.; Wu, W.; Wang, Y.; Alexander, P.B.; Sun, C.; Gong, Z.; Cheng, J.N.; Sun, H.; Guan, Y.; Xia, X.; et al. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. *Nat. Commun.* **2018**, *9*, 5361. [CrossRef] [PubMed]

74. Stencel, K.; Chmielewska, I.; Milanowski, J.; Ramlau, R. Non-small-cell lung cancer: New rare targets—New targeted therapies—state of the art and future directions. *Cancers* **2021**, *13*, 1829. [CrossRef]

75. Schulze, A.B.; Evers, G.; Kerkhoff, A.; Mohr, M.; Schliemann, C.; Berdel, W.E.; Schmidt, L.H. Future options of molecular-targeted therapy in small cell lung cancer. *Cancers* **2019**, *11*, 690. [CrossRef]