IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2471~2483
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2471-2483

Computed tomography imaging radiomics: a novel approach to early-stage non-small cell lung cancer prediction

Raviteja Balekai[1], Mallikarjun S. Holi[2]
[1]Department of Electronics and Communication Engineering, GM Institute of Technology, Visvesvaraya Technological University, Belagavi, India
[2]Department of Electronics and Instrumentation Engineering, University B.D.T College of Engineering, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Nov 19, 2024
Revised Feb 11, 2026
Accepted May 11, 2026

Keywords:

Clinical stage
Computed tomography
Feature selection
Machine learning
Non-small cell lung cancer
Radiomics

ABSTRACT

Radiomics shows promise as non-invasive method for enhancing clinical staging of non-small cell lung cancer (NSCLC) by using quantitative information from computed tomography (CT) scans. This study presents radiomics-based machine learning (ML) approach for staging NSCLC patients into clinical stages I, II, and III based on shape, intensity, and texture features. CT images of 369 NSCLC patients are collected from the cancer imaging archive (TCIA), and extracted 107 radiomic features following image biomarker standardization initiative (IBSI) protocol. The analysis of the sources of variability due to different imaging protocols, using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), showed that these effects were resolved through ComBat harmonization. Recursive feature elimination (RFE) and least absolute shrinkage and selection operator (LASSO) are used for feature selection. Five ML algorithms: logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost) were used, with an 80:20 train-test split and 10-fold cross-validation. The classifier is assessed using accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic (AUROC) curve. The RFE and RF classifier combination performed the best with AUROC of 0.9307 and accuracy of 0.8114. This study illustrates the use of radiomics models in non-invasive classification of NSCLC stages and it is role in clinical decision making.

Corresponding Author:

Raviteja Balekai
Department of Electronics and Communication Engineering, GM Institute of Technology
Visvesvaraya Technological University
Belagavi, India
Email: ravitejj10@gmail.com

1. INTRODUCTION

Lung cancer is the most common cause of cancer-related mortality in the world, and in 2022 alone it was estimated to cause 2.48 million new cases and 1.81 million deaths [1]. The most common subtype form is non-small cell lung cancer (NSCLC), which constitutes approximately 85% of the cases [2], and it includes adenocarcinoma (ADC), squamous cell carcinoma (SCC), and large cell carcinoma (LCC), each with it is own biological behavior and treatment response. Although diagnosis and treatment have advanced, the mortality rate of lung cancer remains high, and the 5-year survival of the NSCLC patients is about 26.9% [3]. Early diagnosis of NSCLC is essential in facilitating curative therapy, which can significantly enhance patient outcomes.

The staging of NSCLC is based on the TNM classification system, which assesses the size of the tumor (T), lymph node (N), and metastasis (M). The stages are I (localized tumor), II (spread to lymph nodes or structures that are nearby), III (spread within the chest), to IV (spread to distant organs). Detecting the disease at the earliest stage can improve survival but since most patients are asymptomatic by the time the disease develops, the currently available treatment approaches are ineffective.

In identifying and diagnosing NSCLC, various imaging techniques are usually used such as, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) [4]. CT scans play the key role in the diagnosis and staging of lung cancer, which is characterized by high-resolution images, which can be used to provide an anatomical understanding of lung tumors. Traditional radiological evaluation of CT images is limited even though it is widely used. The procedure is very much reliant on the experience of the radiologist and tends to concentrate on the visual features of the tumors such as size, shape, and location. Such examinations are subjective and thus causing disparity in diagnosing and staging. In addition, traditional imaging procedures might not be able to identify slight variation between early and late stages of NSCLC decreasing the possibility of early treatment.

In order to address these constraints, radiomics has become an effective method in cancer studies, allowing the identification of quantitative characteristics of medical images, which can be used to detect tumor properties that are unobservable by the human eye [5]. These characteristics, including texture, shape, and intensity, represent a complete tumor phenotype and have demonstrated potential in a wide range of oncological uses, such as tumor characterization, prediction of treatment response, and prognostication of outcome [6]–[8]. Radiomics can be used to identify patterns and correlations that can guide clinical decision-making by applying machine learning (ML) models to such features and help to predict the stage of NSCLC earlier [9]–[11].

This study seeks to build a radiomics model to differentiate clinical stage (I to III) of NSCLC using CT images based on the characteristics of tumors. Stage IV, which is a progressive disease where metastasis occurs, is usually omitted since it has very specific clinical indicators that can be easily detected by conventional imaging. Besides, the treatment of stage IV is normally palliative unlike stage I to III where treatment is more diverse. Therefore, prioritizing stages I to III enables the model to inform decisions in cases where the treatment choices are quite different depending on the stage.

A radiomics approach is present to classify the NSCLC clinical stages using ML from two publicly available multi-institutional datasets. The radiomic features were extracted according to the image biomarker standardization initiative (IBSI) standards [12] to provide consistent and reproducible methodology across all studies. The heterogeneity in the imaging protocols is analyzed and mitigated using ComBat harmonization. least absolute shrinkage and selection operator (LASSO) and recursive feature elimination (RFE) were used to minimize the dimensionality of the large number of radiomic features generated from each image. Finally, synthetic minority oversampling technique (SMOTE) was used to account for the significant class imbalance inherent in the patient population. The multiple ML classifiers were systematically compared to determine which model best identified the correct clinical stage.

The key contributions of this work are: First, develop a robust radiomics pipeline that integrates standardized feature extraction, batch harmonization, and model validation. Second, identify a sub-group of discriminative radiomic features that can derive clinically relevant tumor characteristics of NSCLC stage. Third, give a comparative analysis of ML algorithms, which proves that ensemble-based models are effective in the context of radiomics-based stage classification. The paper is organized as follows: section 2 describes the datasets, feature extraction and selection, model training and validation. Sections 3 discuss the results and compare them with findings from previous studies. Lastly, section 4 concludes this study.

2. METHOD

2.1. Dataset

In this study, the CT images were obtained from two publicly available multi-institutional datasets: dataset 1–the NSCLC-Radiomics dataset [13], and dataset 2–the NSCLC-radiogenomics dataset [14]. Dataset 1 contains CT images from 422 patients, including 51 with ADC, 152 with SCC, 114 with LCC, 63 as not otherwise specified (NOS), and 42 as not available (NA) due to incomplete diagnostic information. Each case includes 3D gross tumor volumes (GTV) delineated by oncologists along with associated clinical outcomes. Following visual inspection using the 3D Slicer tool [15], a selection of 250 patient records with comprehensive clinical data was made for further analysis. The dataset 2 comprises CT images from 144 NSCLC patients, including 112 with ADC, 29 with SCC, and 3 with NOS. These images include segmentations in digital imaging and communications in medicine (DICOM) format and clinical outcomes in CSV files. Initial segmentations were done with a semi-automatic algorithm and later refined by an expert. After verification using the 3D Slicer tool, 119 patient records were chosen for analysis.

The study excluded patients whose clinical stage labels were lacking or ambiguous, had incomplete or corrupted CT images, and tumor segmentation masks were inconsistent or absent to guarantee quality and reproducibility of data. After this process of inclusion and exclusion, a total of 369 patient records were retained from two datasets to continue with the subsequent analyses. In this research, CT scans were acquired from different institutions, therefore, there are variations in the image acquisition protocols due to the differences of the scanner manufacturers and the image acquisition parameters. In order to evaluate this heterogeneity scientifically, DICOM metadata were analyzed with the help of 3D Slicer, the details are provided in Table 1.

Table 1. Imaging parameters of each CT scan

Dataset	Pixel spacing [x, y] in mm	Slice thickness [z] in mm	Matrix size [x,y]
Dataset 1	[0.97656250, 0.9765625]	2.99	[512×512]
Dataset 2	[0.597656, 0.97656]	[0.625–2.5]	[512×512]

2.2. Feature extraction and batch harmonization

The PyRadiomics library [16], integrated into the 3D Slicer, an open-source software for visualization and image analysis, was used to extract radiomic features from the GTV, which was resampled to a resolution of 1×1×1 mm^3. This preprocessing step was carried out to prepare the sample consistent, and the following feature extraction and modeling steps more reliable. The feature extraction was done based on the IBSI feature definitions. 107 radiomic features is extracted for each patient, which were grouped into three categories: i) morphological (14 features), ii) first-order statistical (18 features), and iii) texture (75 features). All the radiomic features were z-score normalized to get a normal distribution.

Radiomic features can be affected by the scanner type, acquisition sequences and reconstruction parameters, particularly in multi-site data. In this study, CT images were acquired from different institutions and different scanner models, and so, batch effect was taken into account prior to building a model [17]. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were performed on standardized features and showed clustering by dataset, suggesting that there was a presence of scanner- and site-specific variations. To eliminate the effects, ComBat harmonization was performed, with dataset origin as a batch covariate. After harmonization, the PCA and t-SNE clustering values were lower and features were better distributed, confirming the successful removal of non-biological variability, while preserving tumor-specific information. This resulted in better and stronger radiomic features for modeling.

2.3. Feature selection

This step is a crucial in selecting radiomic features to predict the stages of NSCLC in CT scans. In this research, the RFE [18] and LASSO [19], [20] methods were employed for feature selection. RFE was used as a wrapper-based feature selection technique. RFE removes features one at a time and fits the model with the remaining features, and the importance of the features is measured according to a specific ML model. This process is repeated until the number of features is equal to the desired number of features to be retained, and the least important features are dropped. Moreover, LASSO uses L1 regularization, which shrinks the absolute values of feature coefficients, thus setting less important features' coefficients to zero. This eliminates non-important or redundant features, and keeps only those that play a role in the predictive model. Through feature reduction, LASSO reduces model complexity, enhances interpretability and reduces overfitting. The expression of Lasso is shown in (1).

$$J(\theta) = \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{m} \theta_j \, x_j^{(i)})^2 + \lambda \sum_{j=0}^{m} |\theta_j| \quad (1)$$

Where, n is the number of observations, m is the number of features, $x^{(i)}$ is ith sample, $y^{(i)}$ is observed output of the ith sample, θ is the regression coefficient, and $\sum_{j=0}^{m} |\theta_j|$ is regularization term with the regularization parameter λ.

By incorporating LASSO and RFE into the feature selection process, this study aims to enhance the performance of the predictive model. It also seeks to improve interpretability by focusing on the most discriminative radiomics features. In addition, this approach helps mitigate the risk of overfitting and computational complexity.

2.4. Model training and validation

This study aimed to create a framework of classification in predicting the clinical stages of NSCLC. There was a significant imbalance in classes in the dataset with stage I (n =118), stage II (n =49), and

stage III (n =202) cases. To reduce the possibility of information leakage, feature selection based on RFE and LASSO and oversampling was only used on the training data in each fold of the cross-validation. The class imbalance was addressed by using the SMOTE [21] to the training dataset, in which the synthetic minority class examples were generated using the k-nearest neighbor algorithm approach, resulting in a better class balance and better performance of the models. Random selection of the subset of training and test data was done with an 80-20 split and the whole series of CT images of each patient was used. The models were trained with five different ML classifiers; logistic regression (LR), decision tree (DT), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost). The evaluation with multiple classifiers is essential in radiomics studies, since classifiers can achieve different performances in terms of features, class imbalance, and complexity of the data association. To evaluate the reliability and robustness of the models, the average performance of accuracy, sensitivity, specificity, F1-score, and the area under the receiver operating characteristic (AUROC) across the folds was calculated.

3. RESULTS AND DISCUSSION

3.1. Results

A total of 369 patients with NSCLC were studied. The average age of the patients was 68.80 years, and 256 (69.37%) of the patients were male. Histological subtypes of the patients were: 135 (ADC), 145 (SCC), and 89 (LCC). Regarding clinical stages, there were 118 patients (31.97%) in stage I, 49 patients (13.27%) in stage II, and 202 patients (54.74%) in stage III. Tables 2 and 3 provides summary outlining demographic and clinical information across two distinct datasets.

Table 2. Demographic and clinical characteristics of the NSCLC-Radiomics dataset

Subject characteristics	ADC	SCC	LCC
Number of subjects	42	119	89
Gender male	26	84	55
Gender female	16	35	34
Age mean age (years)	67.31	69.74	67.32
Age range	45.72-85.60	33.68-88.38	46.34-91.70
Stage I	9	18	13
Stage II	3	20	4
Stage III	30	81	72

Table 3. Demographic and clinical characteristics of the NSCLC-Radiogenomics dataset

Subject characteristics	ADC	SCC
Number of subjects	93	26
Gender male	69	22
Gender female	24	4
Age mean age (years)	68.96	70.69
Age range	43-87	57-83
Stage I	62	16
Stage II	16	6
Stage III	15	4

Each tumor volume underwent extensive characterization by extracting a comprehensive set of 107 features. These quantitative measures can be categorized into three primary groups: i) morphological (shape) attributes, ii) statistical attributes of the first order, and iii) texture attributes. The texture features consist of 24 the gray level co-occurrence matrix (GLCM), 14 gray level dependence matrix (GLDM), 16 the gray level run length matrix (GLRLM), 16 the gray level size zone matrix (GLSZM), and 5 the neighboring gray tone difference matrix (NGTDM) features.

The analysis of standardized radiomic features with PCA and t-SNE was done to determine the presence of batch-related variability. The dimensionality reduction methods were used to visualize the distributions of features and determine possible cluster trends related to the source of data. Figure 1 clearly shows dataset-specific clustering, which is strong evidence of large batch effects of scanners, acquisition protocols, and institutional practice. Besides visual examination, the statistical summaries (in batch) of mean, median, and standard deviation of features were calculated to measure the inter-dataset variation.

To reduce these non-biological variations the ComBat harmonization algorithm was used on the radiomic feature matrix. ComBat is an empirical Bayes approach that is commonly used in radiomics literature to correct batch effect whilst maintaining biologically meaningful signal. In this work, the batch covariate was dataset origin, and the harmonization was achieved with the use of the pycombat

implementation at https://github.com/Jfortin1/ComBatHarmonization. This algorithm was used on the transposed feature matrix to make sure that there was proper adjustment amongst features.

After harmonization, the batch correction was again tested by PCA and t-SNE analysis. Figure 2 shows that dataset specific clustering was significantly less, and there was less overlap in feature distributions between batches. These findings support that ComBat was an effective way of reducing scanner- and site-effects without loss of discriminative information about tumors. Such harmonization contributed to the strength and repeatability of the radiomic features that were to be further modeled.

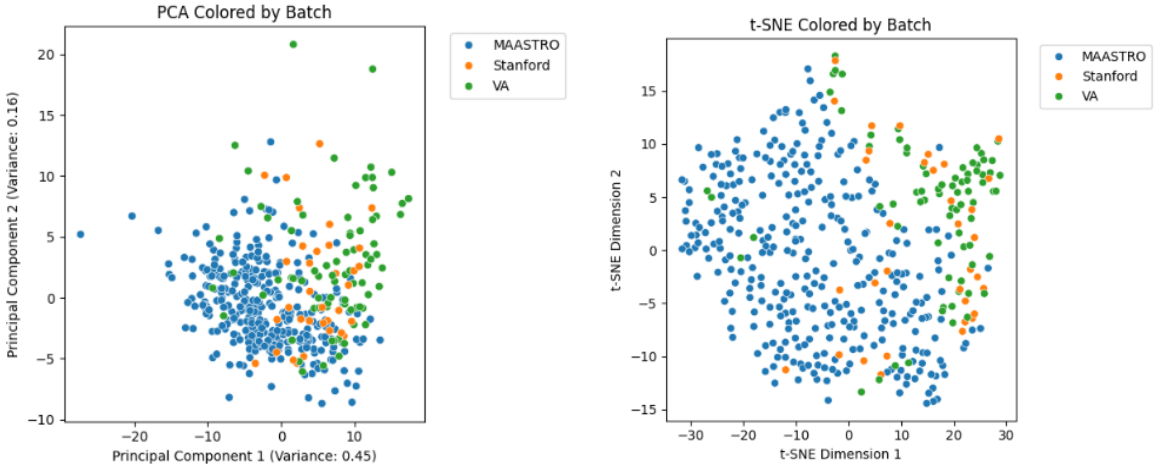

Figure 1. The PCA and t-SNE scatter plot illustrate the distribution of the three stages class dataset

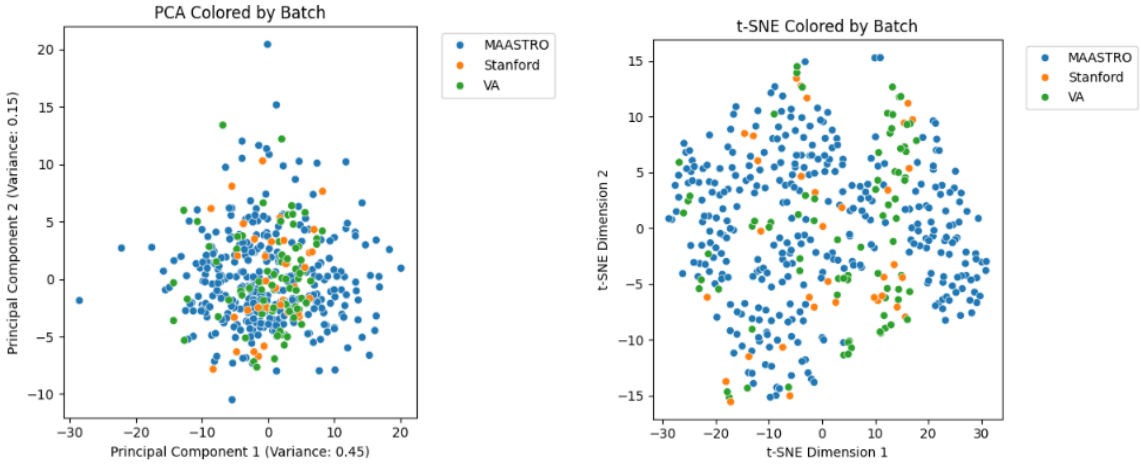

Figure 2. The PCA and t-SNE scatter plot after harmonization

To identify the most distinctive subset of features and enhance model interpretability, \ a feature selection strategy using RFE and LASSO was implemented. For RFE, the feature count was set to 15 while LASSO used a regularization parameter λ of 0.1. The number of features to retain by RFE (n =15) [22], was set based on previous radiomics studies reporting that small feature sets improve generalization while not compromising discrimination in medical classification tasks with multiple classes. The initial experimentation showed that selecting fewer than 10 features led to unstable classification while over 20 features led to increased overfitting without performance gain. Similarly, the LASSO penalty parameter (λ =0.1) [23] was determined empirically on the stability of the features selected across folds and congruence with radiomics-based NSCLC experiments published in the literature. These parameters offer a good compromise between feature selection and model complexity, avoiding overfitting and ensuring both interpretability and high performance.

Some of the radiomic features that are most commonly chosen are clinically relevant. Shape features, including MajorAxisLength, Flatness, and Sphericity, are indicators of tumor size, elongation, and geometrical irregularity, which have been found to rise as disease stage progresses. Imc1, zone entropy, and GrayLevelNonUniformity are texture measures used to measure intratumoral heterogeneity, which has been identified to be associated with aggressive tumor behavior and nodal involvement. The similarity in the characteristics of these features in both feature selection methods indicates that the model is capturing biologically significant features, as opposed to the noise found in the dataset, and enhances clinical interpretability. This selection process with LASSO and RFE highlighted critical features relevant to NSCLC stage classification, as shown in Tables 4 and 5, which detail novel features identified by each method.

Table 4. Selected features for each stage by LASSO algorithm

	Stage 1	Stage 2	Stage 3
Features	Flatness, Major Axis Length, energy, interquartile range, mean, coarseness, strength	90 Percentile, Maximum, Imc1, Imc2, Size Zone Non-Uniformity	Flatness, major axis length, energy, median, root mean squared, Skewness, Imc1, GrayLevelVariance.2, low gray level zone emphasis, zone entropy, coarseness, strength

Table 5. Selected features for each stage by RFE algorithm

	Stage 1	Stage 2	Stage 3
Features	MajorAxisLength, Maximum2DDiameterRow, Maximum3DDiameter, Sphericity, 10Percentile, RootMeanSquared, ClusterProminence, Correlation, Imc1, JointEntropy, SmallDependenceLowGrayLevelEmphasis, RunEntropy, LowGrayLevelZoneEmphasis, SizeZoneNonUniformityNormalized, SmallAreaHighGrayLevelEmphasis	MajorAxisLength, Sphericity, 90Percentile, ClusterProminence, ClusterShade, ClusterTendency, MCC, SumAverage, GrayLevelVariance, HighGrayLevelEmphasis, SmallDependenceHighGrayLevel Emphasis, GrayLevelVariance.2, LowGrayLevelZoneEmphasis, ZoneEntropy, Complexity	Flatness, LeastAxisLength, MajorAxisLength, Mean, RootMeanSquared, Uniformity, Imc1, MCC, DependenceEntropy, DependenceVariance, GrayLevelNonUniformityNormaliz ed GrayLevelNonUniformityNormaliz ed.1, GrayLevelVariance.2, HighGrayLevelZoneEmphasis, SmallAreaLowGrayLevelEmphasis

Tables 6 and 7 shows the results of the five models for staging of NSCLC patients using LASSO and RFE feature selection methods. The area under the curve (AUC) of the LR, SVM, DT, RF, and XGB models using LASSO was 0.7739, 0.7578, 0.6886, 0.8428, and 0.8170. The AUC of the LR, SVM, DT, RF, and XGB models using RFE was 0.8385, 0.8119, 0.7994, 0.9307, and 0.8100.

The RF classifier, combined with RFE feature selection, demonstrated the best performance among the classifiers in our study. The RF model for predicting the stages of NSCLC patients had an accuracy of 0.8114, sensitivity of 0.8135, specificity of 0.8150, F1-score of 0.8142 and an AUC of 0.9307. The confusion matrices and receiver operating characteristic (ROC) plots are shown in Figures 3 to 7 demonstrate a comparison of how well each classification method performed when using feature selection with LASSO; this shows the ensemble models performed better than the linear models at classifying NSCLC stages.

In addition, the confusion matrices and ROC plots obtained from feature selection with RFE are demonstrated in Figures 8 to 12. The results shown in these plots indicate that the RF classifier was able to classify all of the different stages with the best discrimination. This is reflected by its highest AUC.

Table 6. Results of the models for staging of NSCLC patients with LASSO

Classifier	Training accuracy	Test accuracy	Validation accuracy	Sensitivity	Specificity	F1-score	AUC
LR	0.5950	0.5491	0.5890	0.5499	0.5476	0.5488	0.7739
SVM	0.5764	0.5573	0.5725	0.5567	0.5645	0.5606	0.7578
DT	1.0	0.5819	0.5496	0.5859	0.5861	0.5860	0.6886
RF	0.8801	0.6803	0.6388	0.6866	0.6903	0.6884	0.8428
XGB	1.0	0.6393	0.6972	0.6461	0.6428	0.6444	0.8170

Table 7. Results of the models for staging of NSCLC patients with RFE

Classifier	Training accuracy	Test accuracy	Validation accuracy	Sensitivity	Specificity	F1 Score	AUC
LR	0.7231	0.6803	0.6960	0.6798	0.6929	0.6863	0.8385
SVM	0.6053	0.5819	0.5972	0.5746	0.5965	0.5853	0.8119
DT	1.0	0.7295	0.7028	0.7337	0.7328	0.7333	0.7994
RF	1.0	0.8114	0.7706	0.8135	0.8150	0.8142	0.9307
XGB	1.0	0.6065	0.6610	0.6088	0.6019	0.6053	0.8100

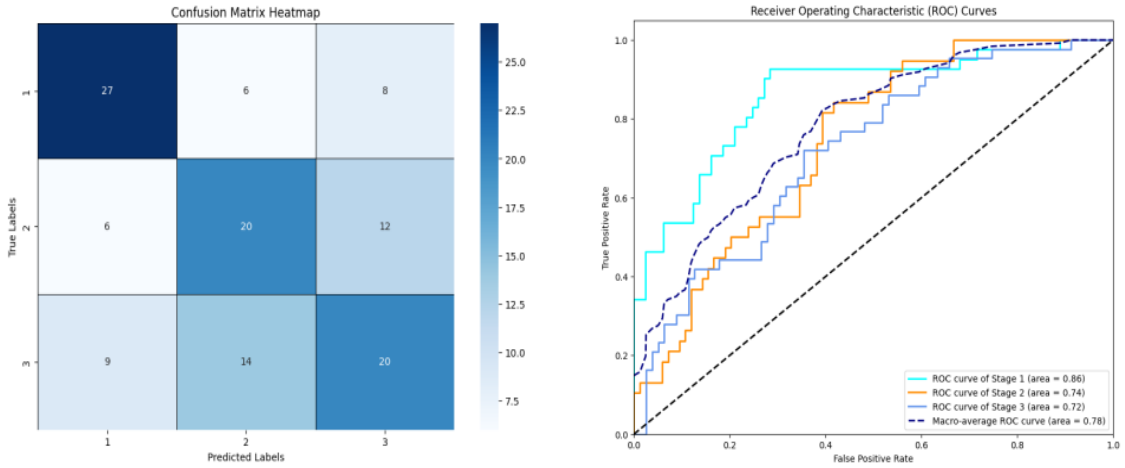

Figure 3. Confusion matrix and ROC curve of LR using LASSO in NSCLC stages classification

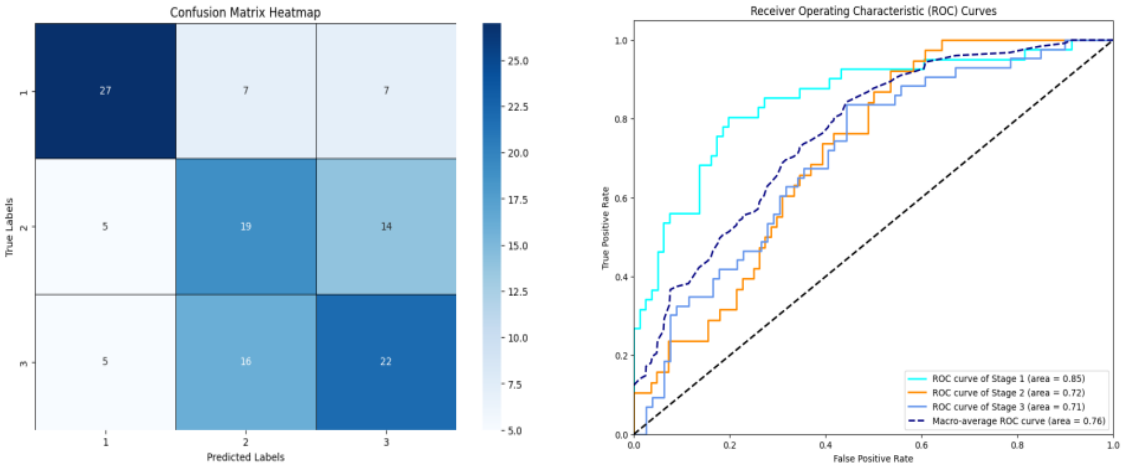

Figure 4. Confusion matrix and ROC curve of SVM using LASSO in NSCLC stages classification

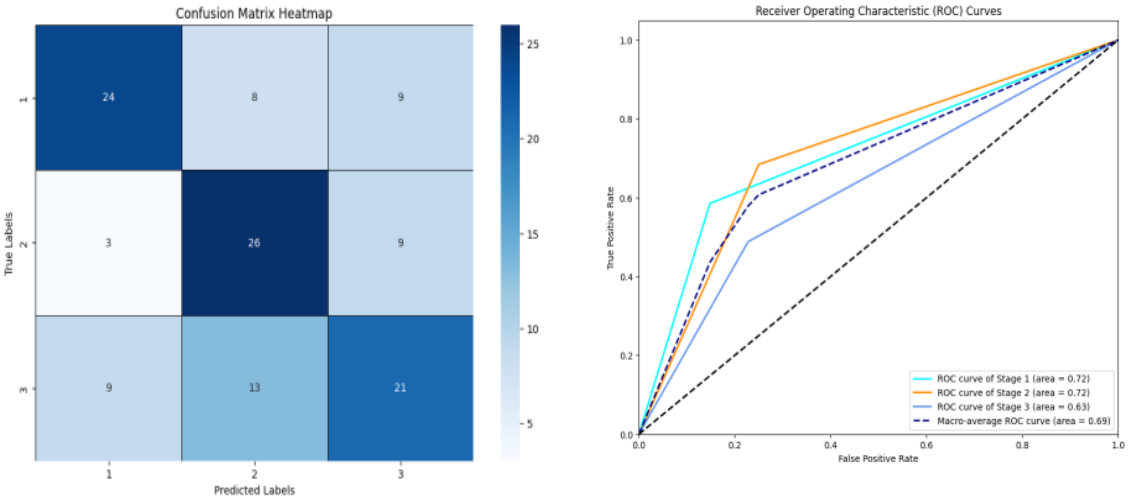

Figure 5. Confusion matrix and ROC curve of DT using LASSO in NSCLC stages classification

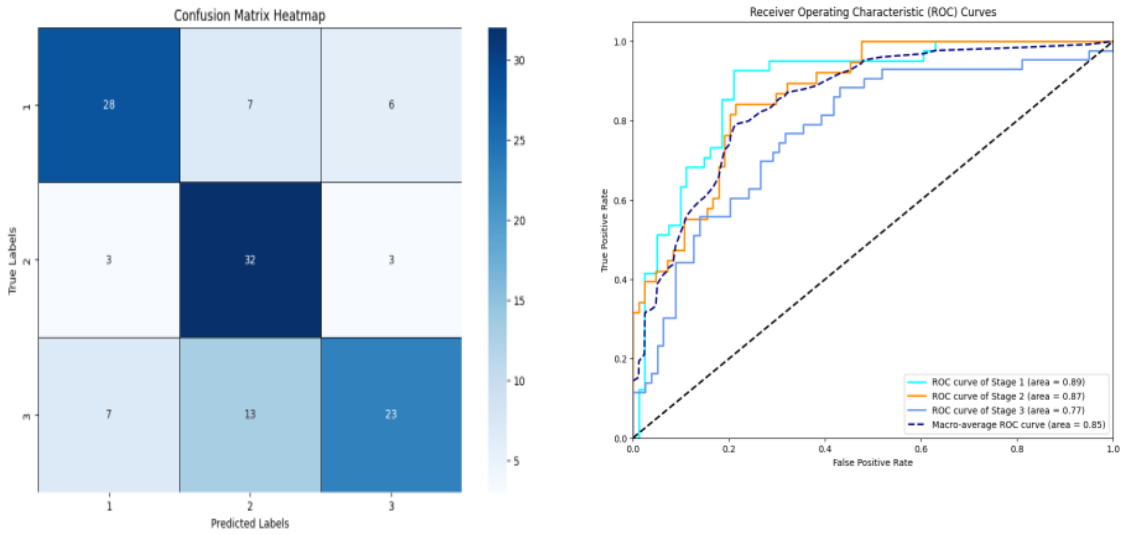

Figure 6. Confusion matrix and ROC curve of RF using LASSO in NSCLC stages classification

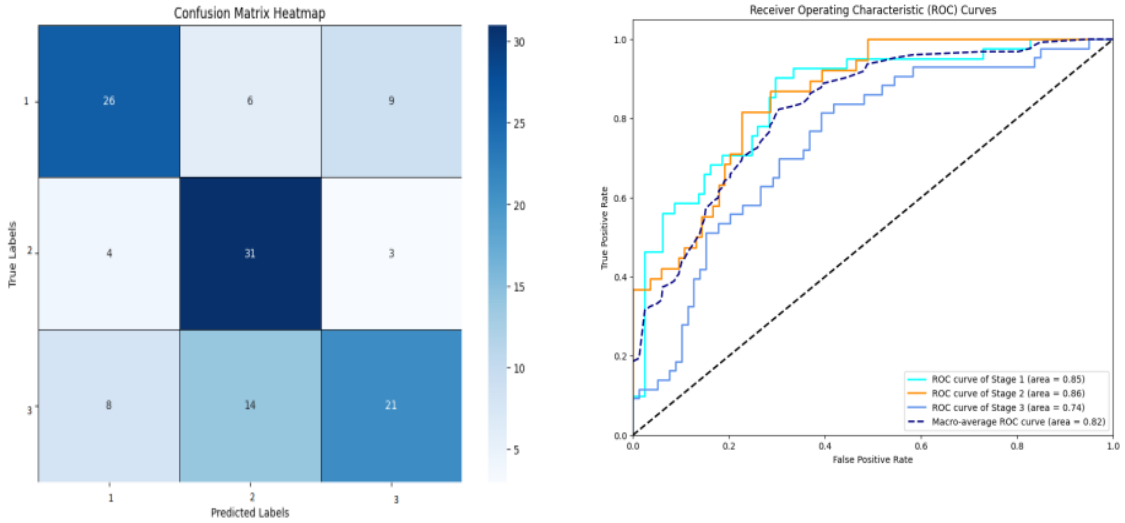

Figure 7. Confusion matrix and ROC curve of XGBoost using LASSO in NSCLC stages classification

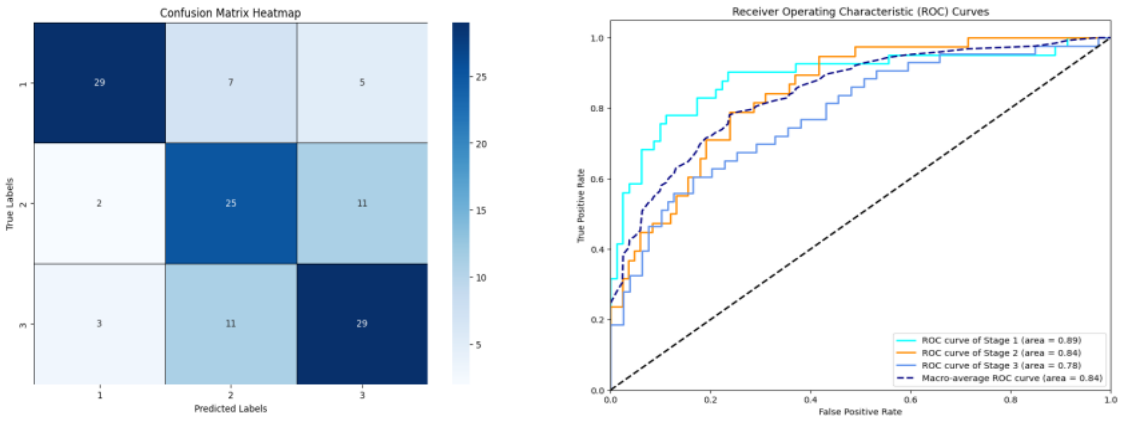

Figure 8. Confusion matrix and ROC curve of LR using RFE in NSCLC stages classification

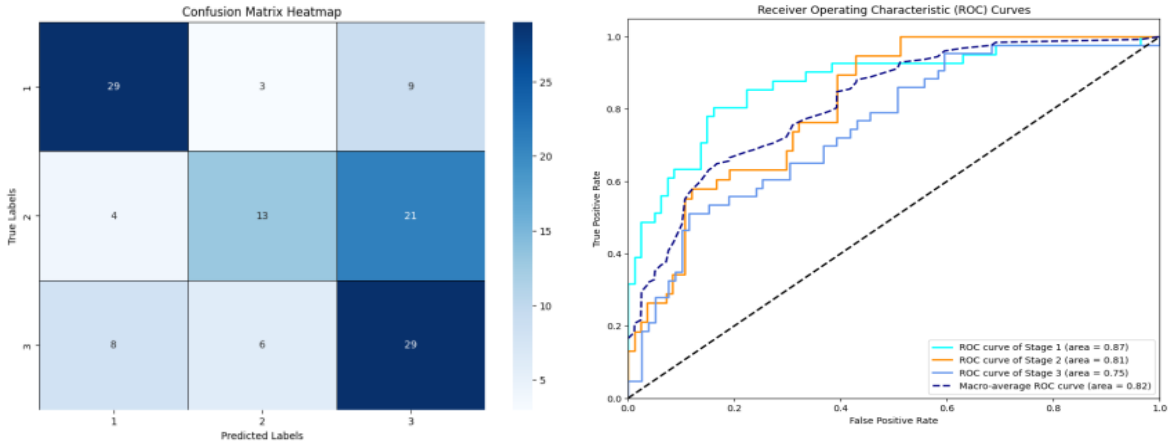

Figure 9. Confusion matrix and ROC curve of SVM using RFE in NSCLC stages classification

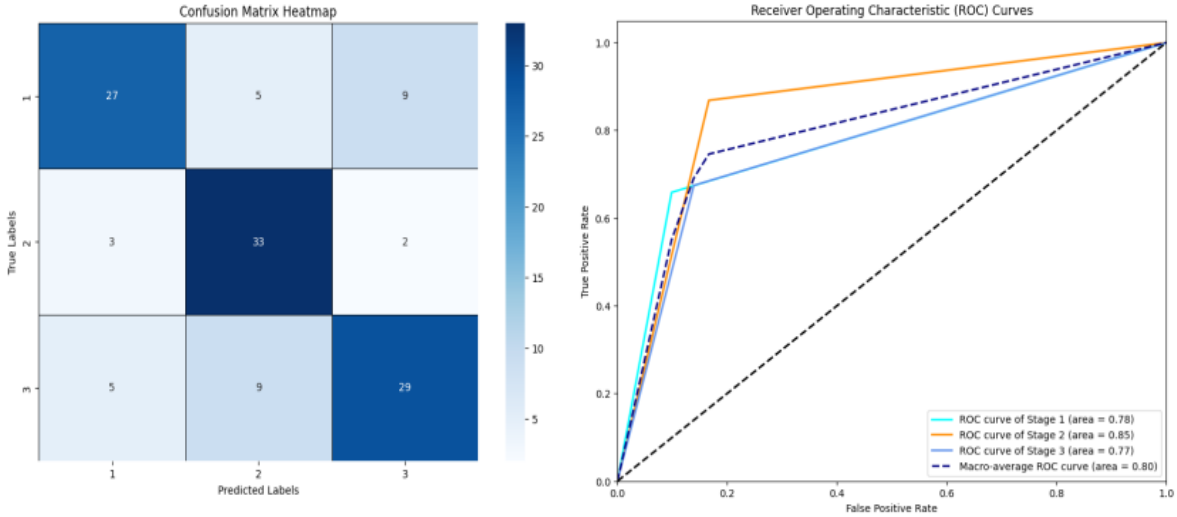

Figure 10. Confusion matrix and ROC curve of DT using RFE in NSCLC stages classification

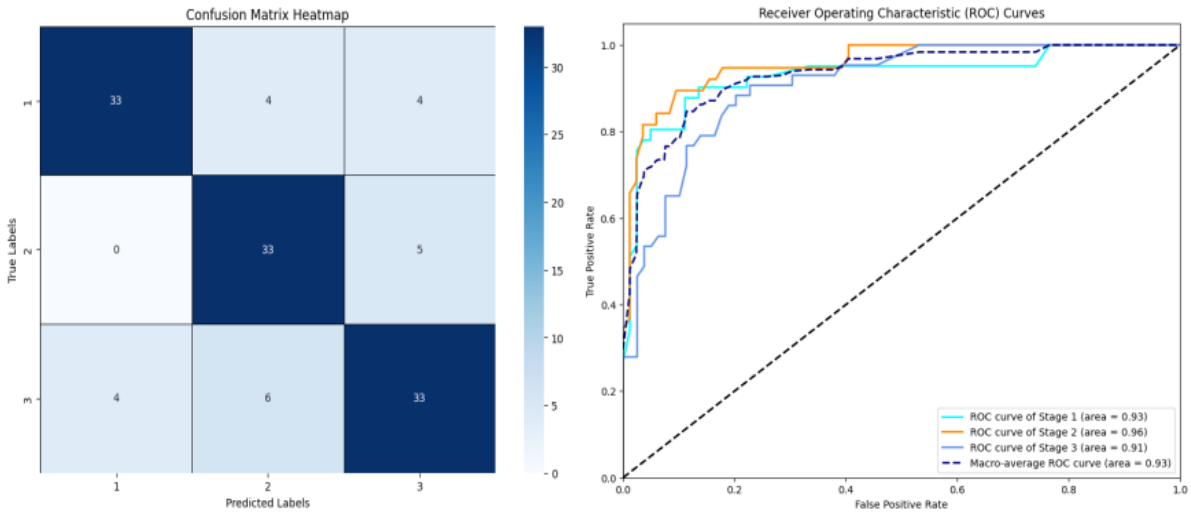

Figure 11. Confusion matrix and ROC curve of RF using RFE in NSCLC stages classification

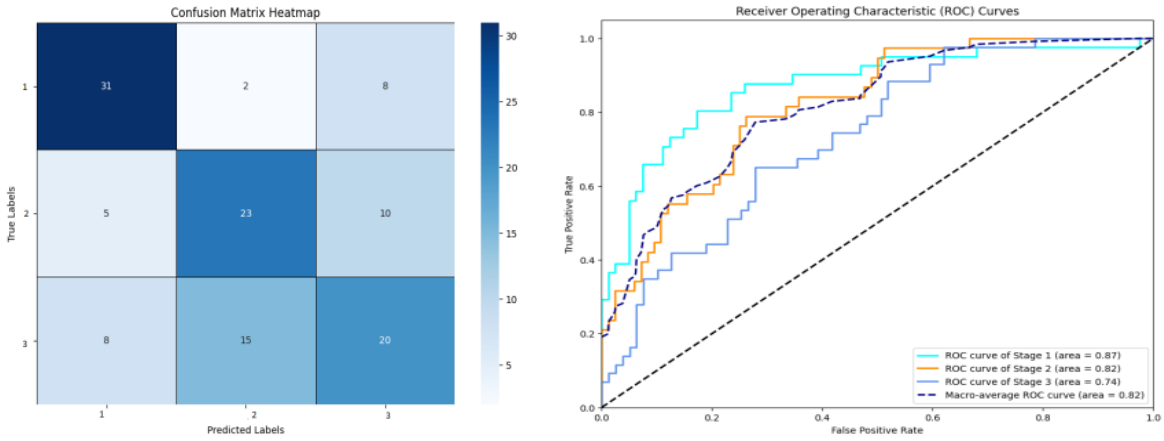

Figure 12. Confusion matrix and ROC curve of XGBoost using RFE in NSCLC stages classification

3.2. Discussion

In recent years, radiomics-based models of NSCLC histological and clinical staging classification have obtained considerable interest owing to their capacity to quantitatively analyze tumor properties in terms of imaging information. The models are a non-invasive, efficient, and objective alternative to the conventional staging procedures like biopsy and manual imaging interpretation. In that our model is built on radiomic features obtained by CT images, which can predict the stage of NSCLC with high accuracy and is consistent with the increasing literature on ML-based cancer diagnostics. Recent investigations have shown that radiomics based on CT and ML can deliver strong performance in terms of NSCLC staging and related clinical prediction activities [24], [25].

Some studies have also pointed out the possibilities of lung cancer staging using radiomics. Yu *et al.* [26] proposed the model to predict the pathological stage (IA grade to IV) of NSCLC based on CT radiomic features, and reported that a RF classifier had better average precision in both lung ADC (AP =0.84) and lung SCC (AP =0.62). Nevertheless, their research indicated that they had limitations associated with the generalizability of the model and overfitting. These findings have been strengthened by more recent multicenter studies that have also underscored the need to have strong validation plans that can guarantee the generalizability of results to heterogeneous datasets [27], [28].

Shimada *et al.* [29] investigated the use of CT-based radiomics with an AI-based approach to predict the pathological lymph node metastasis in early-stage NSCLC with an accuracy of 0.65. Similarly, Yin *et al.* [30] demonstrated that the use of CT radiomics combined with deep learning approaches can be used to predict lymph node metastasis of early-stage lung ADC more accurately, which leads to the growing importance of higher ML techniques in improving staging results. Ubaldi *et al.* [31] used a ML model to predict lung cancer stages (I to II) and histology with limited samples of CT data with AUC values of 0.72±0.04 with the RF and 0.84±0.03 with a linear SVM. Lin *et al.* [32] came up with a model that categorizes both histological subtype and clinical stage (I to III) of NSCLC with a staging AUC of 0.881. Recent studies in [24], [33] have also shown that CT radiomics-based ML classifiers, in particular the RF classifiers, could also be strong and stable in the staging of NSCLC when tested on multicenter datasets.

The model will incorporate some of the limitations that were observed in earlier researches. Feature selection with LASSO and RFE was used as the initial stage of data dimensionality reduction, and the latter is important to increase the likelihood of interpreting the results since RFE only considers the most significant radiomic features to stage the data. The success of LASSO-based feature selection in radiomics has been adequately reported. Pasini *et al.* [34] demonstrated that LASSO was effective in eliminating irrelevant features and enhancing model accuracy to predict histopathological subtypes of NSCLC. On the same note, the results indicate that the addition of LASSO or RFE with ensemble-based classifiers, especially the RF, results in greater staging performance.

The second important improvement of our research is the direct management of the class imbalance that is an inherent problem in medical imaging data in which specific cancer stages are underrepresented. Models that are trained with unbalanced samples will give advantage to majority classes and eventually make biased predictions as other previous radiomics studies have observed [35]. This problem is solved by using the SMOTE to train only the training data, and this solved the problem and the classification was better,

especially with the minority-stage classes. This point is supported by the recent results of Dunn *et al.* [36], who have shown that SMOTE can contribute greatly to the predictive power of radiomics-based cancer classification problems.

Moreover, various ML classifiers were tested, such as DT, RF, SVM, LR, and XGBoost, which provided us with an opportunity to provide a detailed comparison of the performance of the algorithms. The most successful model was the RF classifier as it always had higher accuracy, sensitivity, specificity, and the AUC than the other models. This result is in line with the recent studies that indicate that RF models can be employed especially with high-dimensional radiomic data due to their noise immunity and non-linear complex interactions [37], [38].

While the results were encouraging, there are a number of limitations that need to be considered. The first limitation is that our model is based on only one type of imaging (CT). Recent experiments have shown that the use of multimodal imaging such as PET-CT may be able to significantly improve staging and prognostication performance because anatomical and metabolic information is complementary [39], [40]. The second limitation is the independent-data verification. While 10-fold cross-validation has been used to avoid overfitting and ensure internal validity, external validation is essential to ensure that the model can be used on other types of patients and different imaging equipment. Future research will involve incorporating multimodal radiomics features to enhance accuracy and clinical utility of the model, taking confidence interval and statistical significance testing to better quantify the uncertainty in the performance of the model and variability of inter-institutional and scanner variability which is highlighted by Jha *et al.* [41], using large-scale multicentric validations, normalization of the acquisition parameters, stratified results by different scanners, as well as harmonization-aware learning.

4. CONCLUSION

The current study has shown the potential of radiomic features extracted from CT depicting the ability to predict the clinical stages of NSCLC using standardized feature selection and advanced ML algorithms. Considering the best outcome with the RF predictor the proposed model was able to achieve high levels of accuracy with a combination of IBSI compliant feature extraction, batch effect harmonization, leakage aware feature selection and class imbalance correction in a rigorous validation approach. While these are encouraging results, some limitations are acknowledged. Only a single imaging was considered and external validation was not performed on additional cohorts. These will also be tackled to improve the generalizability and to enable the translation into clinical practice. Future work will be undertaken to validate the proposed approach on multi-center data, to include multi modal imaging (such as PET or MRI) and to add clinical information to further improve the prediction accuracy and clinical interpretability. And will include confidence intervals and hypothesis testing in order to scale up the uncertainty in the model performance. Overall, this paper shows the potential of radiomics of CT as a non-invasive decision support tool for staging of NSCLC. As they are further validated and their methodologies refined, these methods can be used to stratify risks more accurately, create individualized treatment plans and achieve better clinical outcomes in patients with NSCLC.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Raviteja Balekai	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Mallikarjun S. Holi		✓		✓		✓				✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest associated with this publication.

DATA AVAILABILITY

The dataset used in this study is publicly available in references [13], [14].

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/caac.21834.

[2] J. U. Lim, "Overcoming osimertinib resistance in advanced non-small cell lung cancer," *Clinical Oncology*, vol. 33, no. 10, pp. 619–626, Oct. 2021, doi: 10.1016/j.clon.2021.07.015.

[3] A. Cortellini *et al.*, "Determinants of 5-year survival in patients with advanced NSCLC with pd-l1≥50% treated with first-line pembrolizumab outside of clinical trials: results from the pembro-real 5y global registry," *Journal for ImmunoTherapy of Cancer*, vol. 13, no. 2, Feb. 2025, doi: 10.1136/jitc-2024-010674.

[4] M. Wang *et al.*, "An initial study on the comparison of diagnostic performance of ^{18}F-FDG PET/MR and ^{18}F-FDG PET/CT for thoracic staging of non-small cell lung cancer: focus on pleural invasion," *Revista Española de Medicina Nuclear e Imagen Molecular*, vol. 42, no. 1, pp. 16–23, Jan. 2023, doi: 10.1016/j.remnie.2021.12.007.

[5] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[6] M. Selvam, A. Sadanandan, A. Chandrasekharan, S. Ramesh, A. Murali, and G. Krishnamurthi, "Radiomics for differentiating adenocarcinoma and squamous cell carcinoma in non-small cell lung cancer beyond nodule morphology in chest CT," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-83786-6.

[7] Z. Ma *et al.*, "Preoperative ct-based radiomic prognostic index to predict the benefit of postoperative radiotherapy in patients with non-small cell lung cancer: a multicenter study," *Cancer Imaging*, vol. 24, no. 1, May 2024, doi: 10.1186/s40644-024-00707-6.

[8] J. Wang *et al.*, "CT radiomics-based model for predicting TMB and immunotherapy response in non-small cell lung cancer," *BMC Medical Imaging*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12880-024-01221-8.

[9] D. Fotopoulos, D. Filos, E. Xinou, and I. Chouvarda, "Towards lung cancer staging via multipositional radiomics and machine learning," in *16th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2023, pp. 317–324, doi: 10.5220/0011781500003414.

[10] J. Bin *et al.*, "Predicting invasion in early-stage ground-glass opacity pulmonary adenocarcinoma: a radiomics-based machine learning approach," *BMC Medical Imaging*, vol. 24, no. 1, Sep. 2024, doi: 10.1186/s12880-024-01421-2.

[11] Y. Shi *et al.*, "Machine learning prediction models for different stages of non-small cell lung cancer based on tongue and tumor marker: a pilot study," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 2023, doi: 10.1186/s12911-023-02266-5.

[12] A. Zwanenburg *et al.*, "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[13] H. J. Aerts *et al.*, "Data from NSCLC-radiomics," *The Cancer Imaging Archive*, 2015, doi: 10.7937/K9/TCIA.2015.PF0M9REI.

[14] S. Bakr *et al.*, "Data for NSCLC radiogenomics collection," *The Cancer Imaging Archive*, 2017, doi: 10.7937/K9/TCIA.2017.7hs46erv.

[15] A. Fedorov *et al.*, "3D slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[16] J. J. M. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[17] D. Du *et al.*, "Impact of harmonization and oversampling methods on radiomics analysis of multi-center imbalanced datasets: application to PET-based prediction of lung cancer subtypes," *EJNMMI Physics*, vol. 12, no. 1, Apr. 2025, doi: 10.1186/s40658-025-00750-7.

[18] P. Rani, R. Kumar, A. Jain, and S. K. Chawla, "A hybrid approach for feature selection based on genetic algorithm and recursive feature elimination," *International Journal of Information System Modeling and Design*, vol. 12, no. 2, pp. 17–38, Apr. 2021, doi: 10.4018/IJISMD.2021040102.

[19] A. Perniciano, A. Loddo, C. Di Ruberto, and B. Pes, "Insights into radiomics: impact of feature selection and classification," *Multimedia Tools and Applications*, vol. 84, no. 26, pp. 31695–31721, Nov. 2024, doi: 10.1007/s11042-024-20388-4.

[20] B. Peng *et al.*, "Preoperative computed tomography-based tumoral radiomic features prediction for overall survival in resectable non-small cell lung cancer," *Frontiers in Oncology*, vol. 13, May 2023, doi: 10.3389/fonc.2023.1131816.

[21] Z. Zhang and J. Li, "Synthetic minority oversampling technique based on adaptive local mean vectors and improved differential evolution," *IEEE Access*, vol. 10, pp. 74045–74058, 2022, doi: 10.1109/ACCESS.2022.3187699.

[22] F. Deng, L. Zhao, N. Yu, Y. Lin, and L. Zhang, "Union with recursive feature elimination: a feature selection framework to improve the classification performance of multicategory causes of death in colorectal cancer," *Laboratory Investigation*, vol. 104, no. 3, Mar. 2024, doi: 10.1016/j.labinv.2023.100320.

[23] N. Li and W. Chu, "Development and validation of a survival prediction model in elder patients with community-acquired pneumonia: a mimic-population-based study," *BMC Pulmonary Medicine*, vol. 23, no. 1, Jan. 2023, doi: 10.1186/s12890-023-02314-w.

[24] Y. Kang *et al.*, "Computed tomography-based radiomics model for predicting station 4 lymph node metastasis in non-small cell lung cancer," *BMC Medical Imaging*, vol. 25, no. 1, Jun. 2025, doi: 10.1186/s12880-025-01686-1.

[25] Z.-S. Pu *et al.*, "A CT radiomics-based machine learning model for predicting high-grade pathological components in stage IA lung adenocarcinoma: a two-center study," *Clinical Radiology*, vol. 90, Nov. 2025, doi: 10.1016/j.crad.2025.107077.

[26] L. Yu *et al.*, "Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis," *BMC Cancer*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12885-019-5646-9.

[27] M. U. Suleman *et al.*, "Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings," *Annals of Medicine and Surgery*, vol. 87, no. 12, pp. 8803–8811, Dec. 2025, doi: 10.1097/MS9.0000000000004166.

[28] M. K. Das, "Multicenter studies: relevance, design and implementation," *Indian Pediatrics*, vol. 59, no. 7, pp. 571–579, Jul. 2022, doi: 10.1007/s13312-022-2561-y.

[29] Y. Shimada *et al.*, "Artificial intelligence-based radiomics for the prediction of nodal metastasis in early-stage lung cancer," *Scientific Reports*, vol. 13, no. 1, Jan. 2023, doi: 10.1038/s41598-023-28242-7.
[30] X. Yin *et al.*, "CT-based radiomics-deep learning model predicts occult lymph node metastasis in early-stage lung adenocarcinoma patients: a multicenter study," *Chinese Journal of Cancer Research*, vol. 37, no. 1, pp. 12–27, 2025, doi: 10.21147/j.issn.1000-9604.2025.01.02.
[31] L. Ubaldi *et al.*, "Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples," *Physica Medica*, vol. 90, pp. 13–22, Oct. 2021, doi: 10.1016/j.ejmp.2021.08.015.
[32] J. Lin, Y. Yu, X. Zhang, Z. Wang, and S. Li, "Classification of histological types and stages in non-small cell lung cancer using radiomic features based on CT images," *Journal of Digital Imaging*, vol. 36, no. 3, pp. 1029–1037, Feb. 2023, doi: 10.1007/s10278-023-00792-2.
[33] R. Pan, X. Lu, X. Dong, L. Guo, X. Li, and D. Cao, "Predicting pathological staging of non-small cell lung cancer using a multi-task radiomics model integrating intratumoral and peritumoral features," *Oncology Letters*, vol. 30, no. 3, pp. 1–11, Jul. 2025, doi: 10.3892/ol.2025.15177.
[34] G. Pasini, A. Stefano, G. Russo, A. Comelli, F. Marinozzi, and F. Bini, "Phenotyping the histopathological subtypes of non-small-cell lung carcinoma: how beneficial is radiomics?," *Diagnostics*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/diagnostics13061167.
[35] W.-F. Wu *et al.*, "Predicting the T790M mutation in non-small cell lung cancer (NSCLC) using brain metastasis mr radiomics: a study with an imbalanced dataset," *Discover Oncology*, vol. 15, no. 1, Sep. 2024, doi: 10.1007/s12672-024-01333-1.
[36] B. Dunn, M. Pierobon, and Q. Wei, "Automated classification of lung cancer subtypes using deep learning and CT-scan based radiomic analysis," *Bioengineering*, vol. 10, no. 6, Jun. 2023, doi: 10.3390/bioengineering10060690.
[37] J. Huang *et al.*, "Evaluating histological subtypes classification of primary lung cancers on unenhanced computed tomography based on random forest model," *Journal of Healthcare Engineering*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/8964676.
[38] Z. Jie, Z. Yanting, J. Shuqi, A. Jie, Q. Shijun, and C. Huai, "Machine learning prediction models for staging of non-small cell lung cancer patients using radiomics," *CT Theory and Applications*, vol. 34, no. 5, pp. 855–863, 2024.
[39] J. Gao *et al.*, "The predictive value of [^{18}F]FDG PET/CT radiomics combined with clinical features for EGFR mutation status in different clinical staging of lung adenocarcinoma," *EJNMMI Research*, vol. 13, no. 1, Apr. 2023, doi: 10.1186/s13550-023-00977-4.
[40] X. Wang *et al.*, "Multimodal positron emission tomography/computed tomography radiomics combined with a clinical model for preoperative prediction of invasive pulmonary adenocarcinoma in ground-glass nodules," *Academic Radiology*, vol. 32, no. 11, pp. 6929–6942, Nov. 2025, doi: 10.1016/j.acra.2025.07.067.
[41] A. K. Jha *et al.*, "External validation of robust radiomic signature to predict 2-year overall survival in non-small-cell lung cancer," *Journal of Digital Imaging*, vol. 36, no. 6, pp. 2519–2531, Dec. 2023, doi: 10.1007/s10278-023-00835-8.

BIOGRAPHIES OF AUTHORS

Raviteja Balekai received his M.Tech. in Digital Communication Engineering from M S Ramaiah Institute of Technology, Bangalore, India in 2011. He is a Ph.D. student at Visvesvaraya Technological University, Belagavi, India. He is currently assistant professor at Department of Electronics and Communication Engineering, GM Institute of Technology (Affiliated to Visvesvaraya Technological University, Belagavi), Davangere, Karnataka, India. His research focusses on medical image processing. He can be contacted at email: ravitejj10@gmail.com.

Mallikarjun S. Holi received his Ph.D. in Biomedical Engineering from Indian Institute of Technology, Madras, India in 2004. He is currently professor at Department of Electronics and Instrumentation Engineering, University B.D.T. College of Engineering (a constituent college of Visvesvaraya Technological University, Belagavi), Davangere, Karnataka, India. His research focusses on biomedical instrumentation, biomedical signal and image processing, biomechanics, rehabilitation engineering, and process control. He can be contacted at email: drmsholi@gmail.com.