Predicting overall survival of NSCLC patients with clinical, radiomics and deep learning features

Hemalatha Kanakarajan[1***], Jikai Zhou[2***], Aiara Lobo Gomes[3,2], Petros Kalendralis[2], Wenjie Liang[2] , Fariba Tohidinezhad[4,2] , Andre Dekker[2] , Wouter De Baene[1**], Margriet Sitskoorn[1]

1 Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands
2 Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction,Maastricht University Medical Centre, Maastricht, 6228ET,The Netherlands
3 Institute of Molecular Medicine, RWTH Aachen University, Aachen, Germany
4 Department of Radiology & Nuclear Medicine, Erasmus Universiteit Medical Centre, Rotterdam, 3015 GD,The Netherlands

*Hemalatha Kanakarajan and Jikai Zhou are the shared first authors of this paper.
**Correspondence concerning this article should be addressed to Hemalatha Kanakarajan - H.Kanakarajan@tilburguniversity.edu, Jikai Zhou – *jikai.zhou@maastro.nl* and Wouter De Baene - w.debaene@tilburguniversity.edu

Abstract

Background and purpose

Accurate estimation of Overall Survival (OS) in Non-Small Cell Lung Cancer (NSCLC) patients provides critical insights for treatment planning. While previous studies have shown that radiomics or Deep Learning (DL) features improved prediction accuracy, this study aimed to evaluate whether a model that integrates clinical, radiomics, DL, and dosimetric features outperforms other models developed with only a subset of these features.

Materials and methods

We collected pre-treatment lung CT scans and clinical data for 225 NSCLC patients from the Maastro Clinic: 180 for training and 45 for testing. Radiomics features were extracted using the Python radiomics feature extractor, and DL features were obtained using a 3D ResNet model. An ensemble model comprising XGB and NN classifiers was developed using: (1) clinical features only; (2) clinical and radiomics features; (3) clinical and DL features; and (4) clinical, radiomics, and DL features. The performance metrics were evaluated for the test and K-fold cross-validation data sets.

Results

The prediction model utilizing only clinical variables provided an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.64 and a test accuracy of 77.55%. The best performance came from combining clinical, radiomics, and DL features (AUC: 0.84, accuracy: 85.71%). The prediction improvement of this model was statistically significant compared to models trained with clinical features alone or with a combination of clinical and radiomics features.

Conclusion

Integrating radiomics and DL features with clinical characteristics improved the prediction of OS after radiotherapy for NSCLC patients. The increased accuracy of our integrated model enables personalized, risk-based treatment planning, guiding clinicians toward more effective interventions, improved patient outcomes and enhanced quality of life.

Keywords: Lung cancer, Deep Learning, Overall Survival, Radiomics, Radiotherapy

Introduction

Lung cancer is the most commonly diagnosed malignancy worldwide and remains the leading cause of cancer-related mortality. In 2022, approximately 2.5 million new cases and 1.8 million deaths were reported globally [1]. Based on histopathological classification, lung cancer is broadly categorized into Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC), with NSCLC accounting for 85% to 90% of all cases [2]. Approximately one-quarter of NSCLC patients are diagnosed with a locally advanced stage, and the majority of these cases are considered inoperable [3].

Patients with unresectable locally advanced NSCLC display substantial heterogeneity, resulting in marked variability in Overall Survival (OS) [40]. OS is defined as the time from the initiation of treatment to patient death, irrespective of the cause. In the TNM staging system, NSCLC is categorized from stage I

to IV based on tumor size, regional lymph node involvement, and the presence of distant metastases. Stage III disease, commonly referred to as locally advanced NSCLC, represents a heterogeneous group and is the central focus of the present study. This stage is further subdivided into IIIA, IIIB, and IIIC, each associated with distinct prognostic implications. The 5-year OS rates for patients with stage IIIA, IIIB, and IIIC are approximately 36%, 26%, and 13%, respectively [4], with smoking status significantly influencing the prognosis. In a retrospective study of 2,295 NSCLC patients, the median OS was 13.0 months for current smokers, 16.1 months for former smokers, and 29.0 months for never-smokers [45]. Moreover, additional factors such as tumor size, comorbidities, and treatment strategies (including radiotherapy volume, concurrent versus sequential chemoradiotherapy, and use of consolidation immunotherapy) have all been associated with survival differences [46-48].

OS remains a universally accepted and clinically meaningful endpoint in oncology. In NSCLC, where long-term prognosis is often poor, despite recent therapeutic advancements, accurate OS prediction is of paramount importance. First, it enables clinicians to stratify patients more precisely based on prognostic risk, thereby facilitating individualized treatment planning. Second, it supports more efficient allocation of healthcare resources by identifying patients who are most likely to benefit from intensified therapies or enrollment in clinical trials. Third, it offers more reliable prognostic information to patients and their families, aiding in shared decision-making prior to treatment and enhancing psychological readiness for care [49]. These considerations highlight the substantial clinical value of developing machine learning models capable of accurately predicting OS and constitute the primary motivation for this study.

Prior research has explored various machine learning approaches to enhance OS prediction in NSCLC patients. Siah et al. (2019) developed machine-learning models to predict OS in NSCLC patients, achieving a Concordance Index (C-index) of 0.73 (95% CI, 0.72–0.74) [5]. They identified biomarker

status—specifically PD-L1 expression, EGFR mutation, and ALK rearrangement—as the strongest predictor for OS. However, the limited predictive value of single biomarkers emphasized the need for more comprehensive approaches to improve prediction accuracy [5]. Maria et al. (2024) also emphasized the critical need to develop reliable and accurate tools to predict OS before treatment initiation [50], a perspective shared by other researchers [51]. In this context, radiomics and deep learning have proven to be powerful, complementary tools that leverage advanced imaging analysis and artificial intelligence to enhance OS prediction in NSCLC patients through more accurate, non-invasive, and personalized assessments [12, 37, 52, 53, 54].

Radiomics is a research field that extracts quantifiable data, such as patterns, shapes, and intensity levels, from medical images [6]. These data are used to describe and better understand tumors or other irregularities seen in the images, helping doctors and researchers in diagnosis and treatment planning. The radiomics-based approach has been successfully applied for different endpoints in lung cancer, including overall and progression-free survival [7,8,10]. Recent studies have explored radiomics-based approaches for predicting OS in NSCLC. Sun et al. (2018) applied machine learning to radiomic features in 283 patients and achieved a C-index of 0.68 (95% CI, 0.62–0.74) for OS prediction [7]. Liu et al. (2021) developed a CT-based radiomics model for patients receiving nivolumab, reporting an average AUC of 0.61 ± 0.08 for OS and a C-index of ~0.79 [8]. The results of these studies highlight the applicability of radiomics features in predicting OS and subsequently improving the prognostic accuracy of NSCLC patients.

Deep learning, a groundbreaking advancement in artificial intelligence, has transformed numerous fields, including image processing, computer vision, and natural language processing [9]. Deep learning models have also been used to extract imaging features for cancer treatment [24, 25, 26]. In medical imaging, deep learning methods are used to learn the hierarchical representations of the data. The information extracted by the deep learning models from the images can help in predicting

treatment outcomes [10, 19, 20, 21, 22, 27, 28, 29]. Recent studies have explored the application of deep learning in predicting OS for NSCLC patients. Deep learning models using CT scans have shown significant potential for predicting survival and cancer-specific outcomes in NSCLC patients [22]. Hosny et al. (2018) demonstrated that a 3D CNN could predict OS in patients treated with radiotherapy or surgery, achieving AUCs of 0.70 (95% CI, 0.63–0.78) and 0.71 (95% CI, 0.60–0.82), respectively—outperforming hand-crafted radiomics and clinical models [19]. Afshar et al. (2020) developed the deep learning–based radiomics model, reporting a C-index of 0.68 for OS prediction, notably higher than that of traditional radiomics (≈0.51) [20]. Peng et al. (2024) developed a 3D ResNet50-based model using contrast-enhanced CT to estimate response to concurrent chemoradiotherapy in NSCLC, achieving an AUC of 0.84 in the validation cohort and showing a significant association with OS [21]. Similarly, Xu et al. (2019) integrated serial CT scans using a combined CNN–RNN framework in stage III NSCLC patients, achieving an AUC of 0.74 for two-year OS prediction [10]. Collectively, these studies demonstrate that deep learning models can match or exceed the performance of hand-crafted radiomics in predicting the clinical outcomes.

Integrating diverse data sources has also shown promise. Lai et al. (2020) demonstrated that integrating clinical data into deep learning networks helps develop high accuracy OS prediction models for NSCLC patients [11]. Similarly, combining clinical and radiomics features extracted from CT images has improved prediction accuracy [12, 13, 14, 15]. Caii et al. (2024) explored integrating radiomics and deep learning features to predict OS in NSCLC patients treated with immunotherapy [23]. Hou et al., (2024) further demonstrated the significance of dose distribution maps by combining them with deep learning features to predict the Progression-Free Survival (PFS) of NSCLC patients [31].

To the best of our knowledge, no study has yet integrated hand crafted radiomics and deep learning features with clinical features and dose information to develop machine learning algorithms to predict OS for NSCLC patients. A model trained with all these combined features might predict OS with

higher accuracy than models trained with partial combinations of these features. This study aimed to develop a prediction model that combines the radiomics and deep learning features from the images with the clinical and dose information to predict 12-month OS classification for NSCLC patients. Predicting OS with high accuracy provides a more comprehensive understanding of treatment response prior to treatment initiation and could lead to more personalized and effective interventions that improve treatment outcomes, prolong patient survival, and enhance quality of life.

Methods

Data collection

We retrospectively collected clinical data for 296 NSCLC patients from the Maastro Clinic in Maastricht, The Netherlands. All patients were aged 18 or older. This study was approved by the Institutional Review Board of Maastro Clinic (P0617) and Ethics Committee of Maastricht University Medical Centre (METC 2023-0377). All patients underwent definitive radiotherapy (54–66 Gy) between 2015 and 2022. Some patients received concurrent or sequential chemotherapy in addition to radiotherapy, and others received consolidation immunotherapy with durvalumab. Patients with lung surgery or chest radiotherapy history were excluded, as such treatments may alter tumor morphology, density, or imaging texture, potentially affecting the extraction of radiomics and deep learning features. None of the selected patients had any other primary tumor. Although locally advanced NSCLC is conventionally defined as stage III disease, we included both unresectable stage II and stage III patients in this study. This decision was based on the clinical observation that unresectable stage II and III patients often receive comparable treatment regimens, including definitive radiotherapy and systemic therapies [58, 59, 60]. Additionally, including stage II patients allowed for an increased sample size, thereby enhancing the statistical power and generalizability of our predictive model.

For these 296 patients, we collected the planning CT images taken before radiotherapy. CT images were acquired using a Siemens scanner (software version 4.2.7.0). Scanning parameters included

a tube current of 60 mA, a slice thickness of 3 mm, and a reconstruction diameter of 650 mm. Images were reconstructed using the Qr40f convolution kernel. Pixel spacing was 1.27 mm × 1.27 mm, and images were stored in 16-bit format with Hounsfield Unit (HU) scaling (rescale slope = 1, intercept = –1000). The acquisition covered the thoracic region, and all patients were positioned Head-First Supine (HFS). For all patients, the segmentations of the tumors were obtained through manual delineation by expert radiation oncologists from the Maastro Clinic.

From the initial sample of 296 patients, all patients whose OS status was not known at 12 months from radiotherapy were excluded (n = 46). We also excluded 25 patients for whom we had issues in extracting the radiomics features from the Lung gross tumor volume segmentation (either because the segmentation was missing or more than one segmentation was present). This resulted in a total of 225 patients. For patients with more than one planning CT scan, we took the CT scan that was closest to the radiotherapy date.

Clinical features

The list of clinical factors that we collected from the Maastro clinic were age at radiotherapy, gender, BMI, T classification, N classification, NSCLC stage, histological subtype, laterality, location, World Health Organisation Performance Status (WHOPS), smoking status, presence of diabetes, presence of hypertension, presence of cardiac history, presence of cardiac surgery, presence of vascular issues, presence of cerebra vascular issues, presence of respiratory issues, presence of neoplasm, presence of kidney diseases, presence of liver diseases, creatinine level, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) level, CRP (C-reactive protein) level, FEV_1 (Forced Expiratory Volume in 1 second), DLCO (Diffusing Capacity of the Lungs for Carbon Monoxide), type of chemo sequence (concurrent/sequential), durvalumab administered indicator, and the OS indicator at the end of 12 months from beginning of radiotherapy. We extracted the planned radiotherapy dose from the RTDOSE

DICOM files, which contain 3D dose distributions generated by the treatment planning system. To ensure we used the complete planned dose for each patient, we selected only files where the 'DoseSummationType' was set to 'PLAN'—indicating the total dose from the full treatment plan. The pixel values in these files were converted to physical dose values in Gray (Gy). Rather than using the full 3D dose distribution, we derived each patient's total dose from the 3D dose map and added this value to the clinical dataset. This approach ensures accurate dose extraction directly from the source data.

Missing numerical clinical values were imputed using the iterative imputer provided by the Python scikit learn package (https://scikit-learn.org/stable/modules/impute.html). This iterative imputer models each feature with missing values as a function of other features and uses that estimate for imputation. This is done in an iterated round-robin fashion. At each step, a feature column is designated as output y, and the other feature columns are treated as inputs X. A regressor fits on (X, y) for known y. The regressor is then used to predict the missing values of y. This is done for each missing feature in an iterative fashion and is repeated for 10 imputation rounds. One of the limitations of imputing missing values is that when there are too many missing values, the model may not learn a good predictor for that feature. However, the missing clinical values for each of the factors in our clinical data did not exceed 15% of the total number of records, which is well below levels shown to affect model performance. In fact, Junaid et al [55] demonstrated that Multiple Imputation by Chained Equations (a method similar to IterativeImputer) remains effective, yielding results comparable to complete data, with minimal bias and variance distortion even with up to 50% of missing data. We imputed missing categorical values using probabilistic imputation and applied one-hot encoding [32].

Radiomics features

The segment-based radiomics features were extracted from the radiotherapy treatment planning CT scans using the radiomics feature extractor of the Python Pyradiomics package. We

extracted the radiomics features from the Lung gross tumor volume segmentation made by the expert radiation oncologists at Maastro clinic. The seven groups of features extracted from the Region Of Interest (ROI) of the tumor segmentations were shape-based features (14 features), first-order features (18), Gray Level Cooccurrence Matrix (GLCM) features (24), Gray Level Dependence Matrix (GLDM) features (14), Gray Level Run Length Matrix (GLRLM) features (16), Gray Level Size Zone Matrix (GLSZM) features (16), and Neighbouring Gray Tone Difference Matrix (NGTDM) features (5). The resulting 107 radiomics features were considered in this study. The list of radiomics features extracted is listed in the appendix. The definitions of these radiomics features are given in the Pyradiomics documentation (https://pyradiomics.readthedocs.io/en/latest/features.html). The IBSI (Image Biomarker Standardisation Initiative) has standardized the guidelines for extracting the radiomics features from medical images. PyRadiomics has incorporated most of the IBSI recommendations. The minor deviations from this standard are justified in the documentation (https://pyradiomics.readthedocs.io/en/latest/faq.html).

Deep learning features

We extracted deep learning features from the CT scans using a pre-trained 3D ResNet-18 model [41]. The 3D ResNet-18 model is found to have superior performance when compared to other models for classification tasks [41]. The pre-processing steps included normalizing the pixel intensities between 0 and 1 and ensuring all images are in uniform size by cropping or padding. We extracted the deep features from the output of the GlobalAveragePooling3D layer, which is the second-to-last layer in the model. A 3D convolution was applied to the training data. This convolutional layer was designed with a large kernel size of (3, 7, 7), a stride of (1, 2, 2) for down sampling, and padding of (1, 3, 3) to control the spatial dimensions of the output. The 512 deep learning features extracted using this fine-tuned 3D

ResNet model were combined with clinical and radiomics features to form a combined Python dataframe. The complete process of pre-processing and feature extraction is summarized in Figure 1.

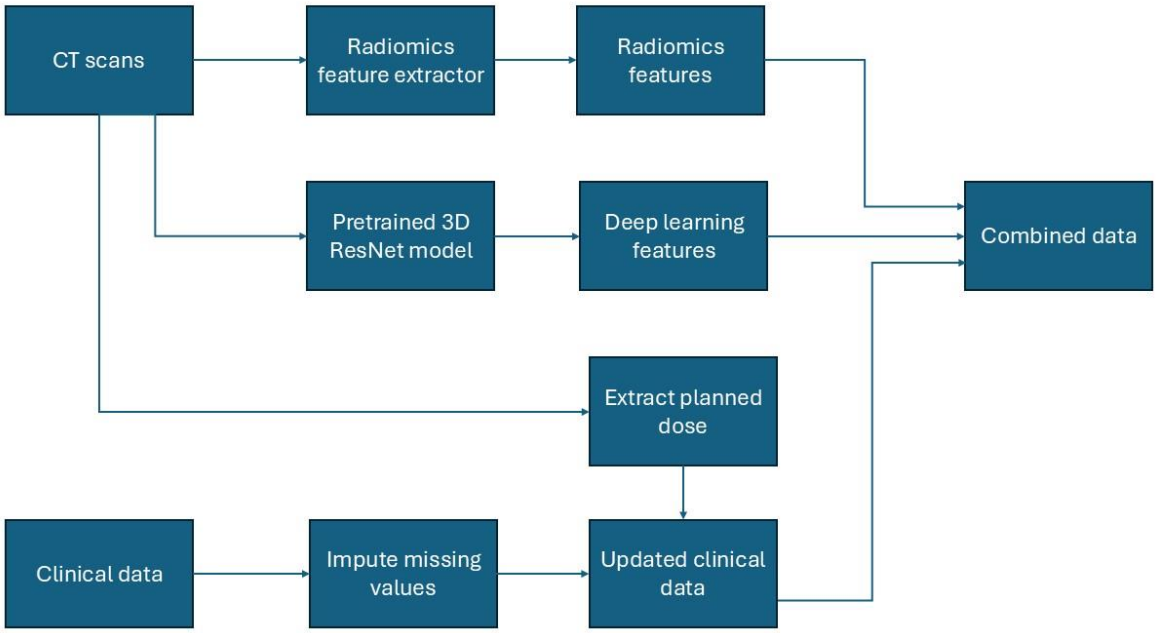

Figure 1: The process of pre-processing, feature extraction, and combining the data.

Model training

We standardized the combined features using the StandardScalar of the Python scikit-learn library (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html). The features with low variance (<0.1) were determined and excluded from the combined dataset to improve prediction accuracy [42]. We applied feature selection using Recursive Feature Elimination (RFE) with a Random Forest Classifier. RFE iteratively removes the least important features and has the potential to improve model performance and reduce overfitting. After RFE, we included only the top 50 most important features. This number is based on the performance of the models in our trial runs with random counts and with the default count (which is half the number of original features). The data set

was then divided into a training data set (80%) and a test data set (20%). The training data set contained 180 patients, and the test data set contained 45 patients. We balanced the training data set using SMOTE (oversampling) [33] and applying random under sampling using Python RandomUnderSampler (https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html). We did oversampling and then undersampling to resolve the class imbalance. The comprehensive analysis of Imani et al. [43] showed that for datasets which are balanced with SMOTE, XGBoost consistently achieved better classification performance when compared to the Random Forest algorithm. Also, the study of Fatima et al. [44] showed that XGBoost consistently achieved better performance for predictive modelling. So, we developed an XGBoost (XGB) classifier for OS classification. We also developed a Neural Network (NN) classifier and compared the performance of the XGB classifier with the NN classifier. We trained the XGB and NN classifiers separately using the training data set. We fine-tuned the hyperparameters of the models to establish the values for the best performance. For XGB, the values are n_estimators=100, learning_rate=0.1, max_depth=4, and scale_pos_weight=1.

We designed our NN classifier using two linear layers with a ReLU (Rectified Linear Unit) activation function in between. ReLU introduces non-linearity into the model by setting all negative values to zero while keeping positive values unchanged, helping the network learn complex patterns more effectively [34, 35]. For training, we used the BCEWithLogitsLoss loss function (https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html), which conveniently combines binary cross-entropy loss with a sigmoid activation in a single step. This loss function simplifies implementation while maintaining accuracy and stability. We optimized the model using Adam, an adaptive learning rate optimizer that adjusts learning rates for individual parameters, enabling faster and more robust convergence. The initial learning rate was set to 0.001. To further improve training stability, we used a learning rate scheduler that decreased the learning rate every five epochs, scaling it

to 80% of its current value. This gradual reduction helps the model make finer adjustments as training progresses, which often improves generalization and final performance. The model was trained for a total of 25 epochs, balancing computational efficiency with sufficient iterations for convergence.

Though XGB classifier achieved better performance for most of the evaluation metrics, NN classifier provided higher Recall when compared to XGB. To harness the advantages provided by both models, we created an ensemble model comprising both XGB classifier and NN classifier. Ensemble models have been shown to outperform individual models in classification tasks in the medical domain [56, 57]. The trained XGB and NN classifiers were combined to form the ensemble model which combines predictions from XGBoost and NN classifiers using a weighted average approach. The predictions are combined with weights of 0.75 for XGB and 0.25 for the NN classifier. The weights are based on the performance of the ensemble model in our trial runs with different weight combinations. We chose the weights which provided the best performance for all the evaluation metrics. The different ensemble models that we created were:

1. Ensemble model with clinical features only.
2. Ensemble model trained with clinical and deep learning features.
3. Ensemble model trained with clinical and radiomics features.
4. Ensemble model trained with clinical, radiomics and deep learning features.

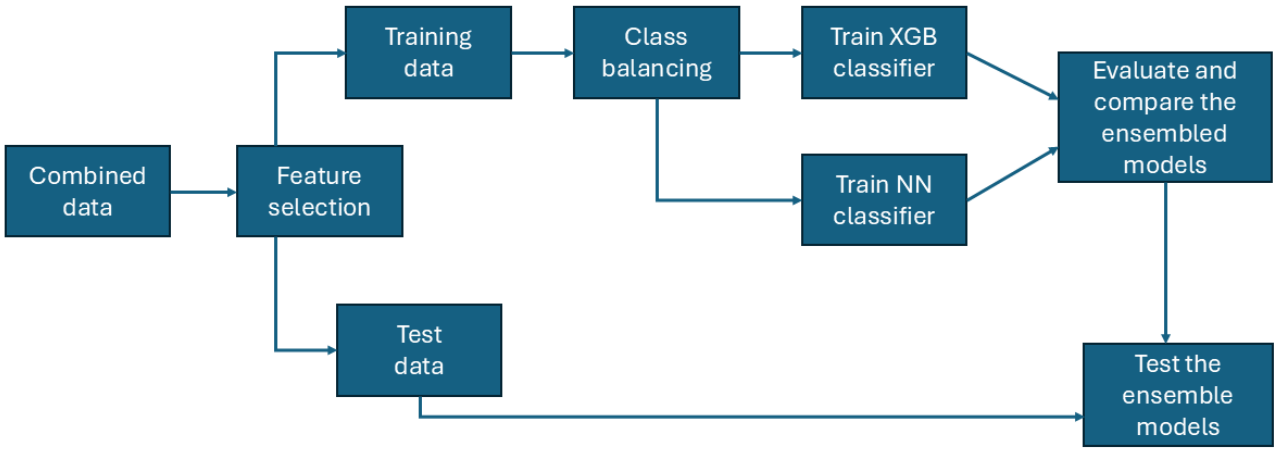

Figure 2: Model training, evaluation and testing.

Model evaluation

The performance of the models was evaluated by measuring the following metrics: classification accuracy, precision, F1 score, recall, Area Under the receiver operating characteristic Curve (AUC), Matthews Correlation Coefficient (MCC), and specificity. The classification accuracy is the ratio of the number of correct classifications to the total number of samples. The precision is the ability of the classifier to identify positive samples correctly, and recall is the ability of the classifier to find all the positive samples. In other words, precision is the ratio of true positive classifications to the total number of positive classifications made by the model, while recall is the ratio of true positive classifications to the total number of actual positives in the dataset. The F1 score gives the balance between precision and recall. A Receiver Operating Characteristic (ROC) is created by plotting the true positive rate vs the false positive rate at various thresholds for defining a positive case. The AUC computes the area under the ROC curve. The MCC is a reliable metric for evaluating the performance of a binary classifier, especially in the presence of imbalanced datasets. The MCC takes into account all four quadrants of a

confusion matrix and provides a balanced measure that considers both the positive and negative classes [36]. Specificity helps understand how well the model performs when predicting the negative class.

A K-fold cross-validation was applied on the model. The different values for K that we used during cross-validation were 5, 7, and 10. The average accuracy and other performance metrics across the different folds were calculated. The best performing model across the different cross validation folds was then tested with a held-out test data set. The accuracy and the other performance metrics were then calculated for the test data set. The pre-processing, feature extraction, model training, and evaluation were performed using Python (version 3.10).

Comparison of models

To evaluate if there was a significant difference in the performance between the four models across the cross-validation data sets and to understand the significance of the performance differences, we statistically compared the performance metrics of the four models. We employed the Friedman test to analyze the significance of the difference in each of the performance metrics across all validation folds across all values of K. When the Friedman test showed a significant difference across the models, a pairwise comparison was done using Wilcoxon signed-rank test with a False Discovery Rate (FDR) correction using the Benjamini-Hochberg method.

Results:

Patient characteristics

A total of 219 NSCLC patients were included in this study, among whom 165 (75.3%) were alive at 12 months and 54 (24.7%) had died. There were no statistically significant differences in most baseline characteristics between the 12-month survival and mortality groups, including age, gender, BMI, tumor stage, and comorbidities (all $p > 0.05$). However, performance status (WHO PS) and history of cardiac surgery showed significant differences. Patients with worse performance status (WHO PS 2–3)

and those with prior cardiac surgery were more frequently observed in the 12-month mortality group ($p < 0.05$). In addition, the use of durvalumab immunotherapy was significantly higher in the survival group (24.2% vs. 5.6%, $p = 0.0026$), suggesting a potential survival benefit. Chemo-radiotherapy sequence showed a trend toward significance ($p = 0.0542$), with concurrent chemoradiation more common in survivors. However, this association should be interpreted with caution, as patients with better performance status were more likely to receive concurrent chemoradiation, indicating possible treatment selection bias. Detailed patient characteristics are summarized in Table 1.

Table 1 shows the characteristics of patients included in our study.

Factor	12-month survival	12-month mortality	P Value*
Age (Year) (mean±SD)	67.34 ± 9.84	69.98 ± 8.67	0.0602
Gender			0.0822
Female	73 (44.2%)	16 (29.6%)	
Male	92 (55.8%)	38 (70.4%)	
BMI	25.93 ± 5.19	26.33 ± 6.05	0.7722
T Stage			0.8749
1	22 (13.3%)	5 (9.3%)	
2	39 (23.6%)	13 (24.1%)	
3	34 (20.6%)	11 (20.4%)	
4	70 (42.4%)	25 (46.3%)	
N Stage			0.8701
0	34 (20.6%)	14 (25.9%)	
1	25 (15.2%)	8 (14.8%)	
2	78 (47.3%)	24 (44.4%)	
3	28 (17.0%)	8 (14.8%)	
Clinical stage			0.3418
IIA	5 (3.0%)	0 (0.0%)	
IIB	24 (14.5%)	7 (13.0%)	
IIIA	61 (37.0%)	27 (50.0%)	
IIIB	58 (35.2%)	14 (25.9%)	
IIIC	17 (10.3%)	6 (11.1%)	
Laterality			0.9188
Left	70 (42.4%)	24 (44.4%)	
Right	95 (57.6%)	30 (55.6%)	
Location			0.6622
Bronchus/Hilum	10 (6.1%)	3 (5.6%)	
Lower	49 (29.7%)	21 (38.9%)	
Middle	7 (4.2%)	2 (3.7%)	

Upper	99 (60.0%)	28 (51.9%)	
WHOPS			<0.05
0	19 (11.5%)	1 (1.9%)	
1	114 (69.1%)	18 (33.3%)	
2	26 (15.8%)	21 (38.9%)	
3	6 (3.6%)	14 (25.9%)	
Smoking			0.6635
Current	45 (27.3%)	12 (22.2%)	
Former	114 (69.1%)	39 (72.2%)	
Never	6 (3.6%)	3 (5.6%)	
Diabetes			0.7098
No	140 (84.8%)	44 (81.5%)	
Yes	25 (15.2%)	10 (18.5%)	
Hypertension			0.2000
No	67 (40.6%)	16 (29.6%)	
Yes	98 (59.4%)	38 (70.4%)	
Cardiac History			1.0000
No	102 (61.8%)	33 (61.1%)	
Yes	63 (38.2%)	21 (38.9%)	
Cardiac Surgery			0.0360
No	137 (83.0%)	37 (68.5%)	
Yes	28 (17.0%)	17 (31.5%)	
Vascular issues			0.0607
No	135 (81.8%)	37 (68.5%)	
Yes	30 (18.2%)	17 (31.5%)	
Cerebro vascular issues			0.6488
No	135 (81.8%)	42 (77.8%)	
Yes	30 (18.2%)	12 (22.2%)	
Respiratory issues			0.7781
No	79 (47.9%)	24 (44.4%)	
Yes	86 (52.1%)	30 (55.6%)	
Neoplasm			0.7990
No	133 (80.6%)	42 (77.8%)	
Yes	32 (19.4%)	12 (22.2%)	
Kidney diseases			0.1571
No	146 (88.5%)	43 (79.6%)	
Yes	19 (11.5%)	11 (20.4%)	
Liver diseases			0.1994
No	159 (96.4%)	49 (90.7%)	
Yes	6 (3.6%)	5 (9.3%)	
Creatinine	83.84 ± 32.26	97.37 ± 85.59	0.3974
CKD-EPI	74.50 ± 27.39	71.39 ± 28.42	0.3863
CRP	39.06 ± 59.48	35.43 ± 44.11	0.3789
FEV1	73.47 ± 21.19	70.48 ± 20.94	0.3908
DLCO	63.65 ± 17.64	59.29 ± 20.38	0.2067

Chemo Sequence			0.0542
Concurrent	101 (61.2%)	23 (42.6%)	
Sequential	27 (16.4%)	14 (25.9%)	
Without chemotherapy	37 (22.4%)	17 (31.5%)	
Durvalumab immunotherapy			0.0026
No	125 (75.8%)	51 (94.4%)	
Yes	40 (24.2%)	3 (5.6%)	
Prescribed total dose(Gy)	61.70 ± 2.85	61.62 ± 2.54	0.9176

Abbreviations: SD = standard deviation; BMI = body mass index; CKD-EPI = Chronic Kidney Disease Epidemiology Collaboration equation; CRP = C-reactive protein; FEV_1 = forced expiratory volume in 1 second; DLCO = diffusing capacity of the lung for carbon monoxide.

*P-values were calculated using the Chi-square test or Fisher's exact test for categorical variables, and Welch's t-test or Mann–Whitney U test for continuous variables, as appropriate. We added the P-values to this table to show the statistical comparison between the two groups on every factor.

The average performance metrics for the four ensemble models across the cross-validation datasets are shown in Table 2. In bold is the value of the best score for the corresponding performance metric.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC	MCC	Specificity (%)
Clinical features only	78.65 (75.91, 81.4)	71.64 (66.5, 76.77)	80.07 (77.27, 82.87)	74.74 (71.47, 78.00)	0.79 (0.77, 0.82)	0.57 (0.52, 0.62)	78.30 (74.95, 81.65)
Clinical and deep learning features	81.23 (76.87, 85.58)	74.05 (68.75, 79.36)	85.55 (80.04, 91.06)	78.75 (74.31, 83.18)	0.82 (0.78, 0.87)	0.63 (0.54, 0.72)	79.23 (74.42, 84.14)
Clinical and radiomics features	77.67 (75.34, 80.00)	68.95 (65.32, 72.58)	80.00 (76.06, 84.52)	73.55 (70.78, 76.32)	0.78 (0.76. 0.80)	0.55 (0.51, 0.60)	76.16 (73.41, 78.91)
Clinical, radiomics and deep learning features	**82.61 (80.10, 85.12)**	**74.75 (70.86, 78.83)**	**86.62 (82.44, 90.78)**	**79.62 (76.96, 82.23)**	**0.84 (0.81, 0.87)**	**0.65 (0.60, 0.71)**	**80.84 (78.09, 83.60)**

Table 2: Average performance of the models along with their 95% confidence intervals on the validation datasets.

The accuracy of the model trained with clinical features only was 78.65%. The model trained with the combination of clinical and deep learning features had an accuracy of 81.23%. The accuracy of the model trained with the combination of clinical and radiomics features was 77.67%. However, the model with the highest prediction accuracy of 82.61% was the model trained with the combination of clinical, radiomics, and deep learning features. This model also achieved the highest precision of 74.75%, Recall of 86.62%, F1 score of 79.62%, MCC of 0.65, specificity of 80.84% and the best AUC of 0.84. The model trained with the combination of all the features had the best score for all performance metrics.

The performance metrics for the four ensemble models for the test data set are shown in Table 3. In bold is the value of the best score for the corresponding performance metric. The performance metrics of the separate XGB and NN classifiers for the test data set are added to the Appendix.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC	MCC	Specificity (%)
Clinical features only	77.55	44.44	40.00	42.11	0.64	0.28	87.18
Clinical and deep learning features	73.47	33.33	30.00	31.58	0.57	0.15	84.62
Clinical and radiomics features	75.51	40.00	40.00	40.00	0.62	0.25	84.62
Clinical, radiomics and deep learning features	**85.71**	**61.54**	**80.00**	**69.57**	**0.84**	**0.61**	**87.18**

Table 3: Performance of the models for the test dataset.

Similar to the performance metrics for the cross validation data set, the model trained with the combination of all the features had the best score for all performance metrics for the test data set. The ROC curve of the model trained with the combination of clinical, radiomics, and deep learning features is shown in Figure 3. The ROC curves of the other three models were added to the Appendix.

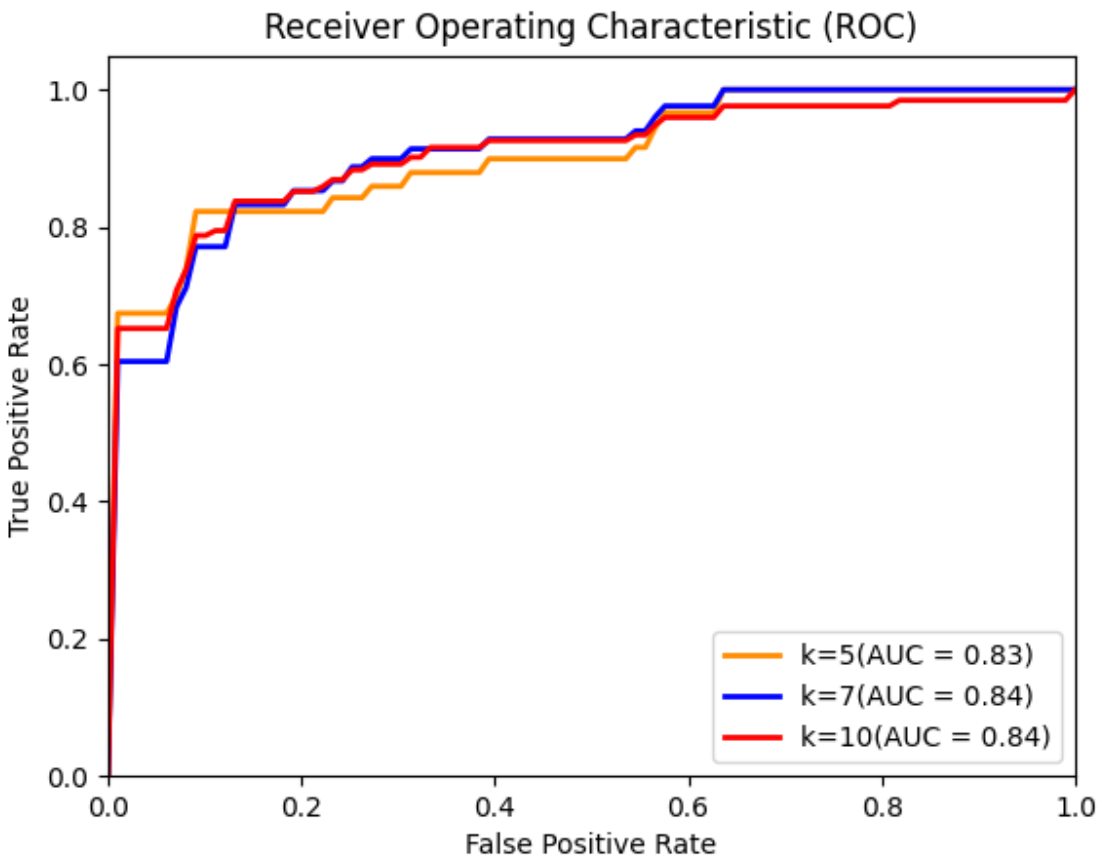

Figure 3: ROC curve of model trained with clinical, deep learning and radiomics features.

The top 10 factors for predicting the OS for each model were added to the appendix. For the combined model, the top 10 features associated with the prediction of the OS were WHOPS, DL_202, DL_301, DL_144, DL_418, DL_25, DL_20, DL_87, DL_229 and DL_133. Except for WHOPS, the rest of these top 10 features are deep learning features. There were no radiomics features in this list. The top 50 features of the combined model included WHOPS, one radiomics feature (original_shape_MajorAxisLength) and 48 deep learning features. WHOPS emerged as the variable with the highest importance in the combined model. Also, in the other three models, WHOPS emerged as the variable with the highest importance.

Outcome of statistical analysis

We used the Friedman test to analyze the significance of the difference in performance metrics across validation folds across all values of K. When the results of this test indicate a statistically significant difference for the metric, we performed pairwise Wilcoxon signed-rank tests with corrections for multiple comparisons using the FDR method to determine which models had statistically different performances.

The results of the Friedman test indicate a statistically significant difference in the accuracy of the four models with a Chi-squared value of 15.25 and a p-value of 0.0016. This means that there is a statistically significant difference in accuracy for at least one of the models. The outcomes of the pairwise Wilcoxon signed-rank tests are presented in Table 4. There is a statistically significant difference (after FDR correction) between the model trained with clinical features only and the model trained with the combination of all features (p-value = 0.015). Also, there is a statistically significant difference between the model trained with clinical and radiomics features and the model trained with the combination of all features (p-value = 0.004). The Friedman test also indicated statistically significant differences for Precision (Chi-squared: 9.44, p-value: 0.0239), Recall (Chi-squared: 8.94, p-value: 0.0299), F1 score (Chi-squared: 16.85, p-value: 0.0007), AUC (Chi-squared: 17.68, p-value: 0.0005) and MCC (Chi-squared: 17.68, p-value: 0.0005). The Friedman test did not indicate a statistically significant difference for Specificity (Chi-squared: 3.74, p-value: 0.2903). The outcomes of the pairwise Wilcoxon signed-rank tests for Precision, Recall, F1 score, AUC and MCC are added to the Appendix. Similar to accuracy, there is a statistically significant difference (after FDR correction) between the model trained with clinical features only and the model trained with the combination of all features for Recall, F1 score, AUC and MCC. There are also statistically significant differences between the model trained with the combination of clinical and radiomics features and the model trained with the combination of all features for F1 score, AUC and MCC.

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	0.115	0.172
clinical features only vs clinical and radiomics features	0.273	0.328
clinical features only vs combination of all 3 features	**0.005**	**0.015**
clinical and deep learning features vs clinical and radiomics features	0.081	0.163
clinical and deep learning features vs combination of all 3 features	0.503	0.503
clinical and radiomics features vs combination of all 3 features	**0.001**	**0.004**

Table 4: Outcome of statistical analysis for accuracy. In bold are the statistically significant values.

Discussion

Accurate prediction of 12-month OS in patients with NSCLC enables more personalized and effective therapeutic interventions, ultimately leading to improved treatment outcomes. In this study, we developed and evaluated four distinct machine learning models using various combinations of clinical features (including radiation dose), hand-crafted radiomics features, and deep learning features to predict the 12-month OS in NSCLC patients treated with radiotherapy. The models included: (1) clinical features only; (2) combination of clinical and deep learning features; (3) combination of clinical and radiomics features; and (4) combination of clinical, radiomics, and deep learning features.

Among the four models evaluated, the model incorporating all three feature types (clinical, radiomics, and deep learning features) achieved the highest predictive performance, with an AUC of 0.84 and an accuracy of 85.71%. This model significantly outperformed the model trained solely on

clinical features and the model trained with the combination of clinical and radiomics features. Notably, the addition of deep learning features or radiomics alone to the clinical features did not significantly improve the prediction accuracy. This suggests that radiomics and deep learning features may offer complementary prognostic information and that their integration with clinical data yields a synergistic effect, enhancing the model's predictive capability.

Our findings align with and expand the conclusions of previous studies. Braghetto et al. (2022) combined clinical, hand-crafted radiomics and deep learning features to predict OS for NSCLC patients treated with radiotherapy or chemoradiotherapy [37]. Their study showed that the models combining clinical, radiomics, and deep learning features provided an AUC of 0.67 compared to an AUC of 0.59 using only clinical features. The study of Liao et al. (2024) showed that the combined model achieved an AUC of 0.80 in predicting PFS for NSCLC patients treated with immunotherapy [38]. Our model trained with clinical, deep learning, and radiomics features provided an AUC of 0.84 for predicting OS in NSCLC patients treated with radiotherapy, indicating a potential performance gain. This suggests that methodological enhancements such as applying SMOTE, imputing missing values, and using an ensemble model may contribute to an increase in the prediction performance. Our previous work on the prediction of local control of brain metastases also demonstrated that models integrating clinical, radiomics, and deep learning features outperform those using only a subset of these features [30], suggesting robustness and generalizability of this integrative modeling approach.

Feature importance analysis revealed that WHOPS, a scale used to assess a cancer patient's general well-being and ability to perform daily activities, was the most important feature in our models, underscoring its known prognostic value of performance status in NSCLC patients. This aligns with the findings from the clinical study of Kawaguchi et al. [39], who identified performance status as an independent prognostic factor for OS in NSCLC. The prognostic strength of WHOPS may be attributed to

the fact that the WHOPS reflects the patient's baseline physical condition, which can influence treatment intensity and decision-making which also carries the risk of confounding our analysis. Differences in baseline status may partially explain the observed variation in survival outcomes. Patients with better performance status and functional capacity tend to have less difficulty in tolerating rigorous cancer treatments and have more favorable outcomes than patients with a worse performance status [61]. Interestingly, the top 10 most predictive features in the combined model included WHOPS and nine deep learning-derived features, but no radiomics features. The top 50 features included WHOPS, one radiomics feature (original_shape_MajorAxisLength) and 48 deep learning features. The planned treatment dose extracted from the RTDOSE DICOM files was also not part of the top 50 features. This could be due to the narrow dose range (54–66 Gy), resulting in low variability across patients and thus offering limited discriminatory power. The higher performance of the combined model when compared to the model trained with clinical and deep learning features suggest that original_shape_MajorAxisLength contains a distinct information that is not captured by clinical or deep learning features. The original_shape_MajorAxisLength feature indicates the degree of elongation of the segmented tumor volume, which can be suggestive of a locally advanced disease. This feature may synergistically interact with the deep learning features and shift decision boundaries of the model in meaningful ways. While the improvement in performance of the combined model was modest when compared to the clinical and deep learning model for the cross validation data sets, the radiomics feature had a greater impact on the performance on the hold-out test set. This shows that the addition of radiomics features to the model improves its robustness and generalization. In general, results suggest that deep learning features may capture more nuanced prognostic information from imaging data than hand-crafted radiomic features. This is in contrast with some other studies which found that radiomics features are more predictive than deep learning features [30, 37]. Further research is needed to evaluate the conditions under which each feature type provides superior prognostic value.

Clinically, these findings have important implications. For patients with unresectable, locally advanced NSCLC, reliably estimating the 12-month OS prior to radiotherapy offers clinicians individualized prognostic information. Such prediction can guide clinical decision-making based on patient-specific risk profiles, enabling the exploration of alternative treatment strategies—for example, implementing more aggressive chemotherapy, immunotherapy, or consolidation therapies in patients with poor predicted prognosis, or considering novel treatment regimens on top of standard care. Personalized OS prediction before treatment provides a valuable resource for clinicians and patients, supporting shared decision-making and developing personalized treatment plans based on each patient's unique needs and risk profile.

Although the primary focus of this study was creating a model for predicting the 12-month OS after radiotherapy, the modeling framework can be adapted to other clinical endpoints such as progression-free survival or to different treatment modalities including chemoradiotherapy and chemotherapy. Similarly, this approach can be extended to predicting the OS at other time points, broadening its potential clinical utility.

While our results are promising, this study has limitations. One limitation of this study lies in the tumor segmentation procedure. Since this study included only patients who received definitive radiotherapy, the tumor segmentations were manually delineated by expert radiation oncologists at Maastro clinic, ensuring high precision in defining the ROI. However, this approach also limits the scalability and applicability of the model. Implementing fully automated deep learning–based segmentation methods may improve both efficiency and reproducibility in future studies. Furthermore, for a more rigorous evaluation of the efficacy and versatility of the models, further investigations involving larger patient cohorts with multi-institutional data are recommended. Our results are from a small cohort of radiotherapy patients from a single institution. An external validation dataset could

significantly improve the generalizability of the OS prediction model and enrich confidence in its broader applicability across diverse patient populations and healthcare domains. Also, because our analysis is based on historical data the model only provides prognostic information and cannot be used to support treatment selection such as between concurrent or sequential radiotherapy and other therapies.

In conclusion, our results demonstrate that integrating clinical, radiomics, and deep learning features substantially improves OS prediction in NSCLC patients treated with radiotherapy. The developed model provides a valuable tool for risk stratification, treatment planning, and patient counseling, paving the way for more personalized and more effective cancer care.

Acknowledgements:

We would like to thank Dr. Dirk De Ruysscher (Department of Radiation Oncology, Maastro) and Dr. Lizza Hendriks (Department of Pulmonary Diseases, Maastricht University Medical Centre+) for their valuable contribution in providing clinical data for this study.

Data Availability:

The data used for this study is available at Maastro and is accessible upon request after necessary approvals.

Funding:

This research is supported by KWF Kankerbestrijding and NWO Domain AES, as part of their joint strategic research programme: Technology for Oncology IL. The collaboration project is co-funded by the PPP Allowance made available by Health Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Consent for publication:

Not applicable.

Contributions:

All authors contributed to this research.

Competing interests

All authors declare that they have no competing interests.

Ethics approval

This study is part of the AI in Medical Imaging for novel Cancer User Support (AMICUS) project. This study was approved by the Institutional Review Board of Maastro Clinic (P0617) and Ethics Committee of Maastricht University Medical Centre (METC 2023-0377). This study adhered to the ethical principles outlined in the Declaration of Helsinki.

Clinical trial number:

Not applicable.

List of abbreviations

OS	Overall Survival
NSCLC	Non-Small Cell Lung Cancer
DL	Deep Learning
AUC	Area Under the Receiver Operating Characteristic Curve
SCLC	Small Cell Lung Cancer
PFS	Progression-Free Survival
HU	Hounsfield Unit
HFS	Head-First Supine
WHOPS	World Health Organisation Performance Status
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
CRP	C-reactive protein
FEV_1	Forced Expiratory Volume in 1 second
DLCO	Diffusing Capacity of the Lungs for Carbon Monoxide
Gy	Gray
ROI	Region Of Interest
GLCM	Gray Level Cooccurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix

GLSZM	Gray Level Size Zone Matrix
NGTDM	Neighbouring Gray Tone Difference Matrix
RFE	Recursive Feature Elimination
XGB	XGBoost
NN	Neural Network
ReLU	Rectified Linear Unit
MCC	Matthews Correlation Coefficient
ROC	Receiver Operating Characteristic
FDR	False Discovery Rate

References:

1.Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 Apr 4. doi:10.3322/caac.21834.

2.Bourreau C, Treps L, Faure S, Fradin D, Clere N. Therapeutic strategies for non-small cell lung cancer: experimental models and emerging biomarkers to monitor drug efficacies. Pharmacol Ther. 2023;242:108347. doi:10.1016/j.pharmthera.2023.108347.

3.Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol. 2016;11(1):39–51. doi:10.1016/j.jtho.2015.09.009.

4. Petrella F, Rizzo S, Attili I, Passaro A, Zilli T, Martucci F, Bonomo L, Del Grande F, Casiraghi M, De Marinis F, Spaggiari L. Stage III Non-Small-Cell Lung Cancer: An Overview of Treatment Options. Curr Oncol. 2023;30(3):3160–3175. doi:10.3390/curroncol30030239.

5.Siah KW, Khozin S, Wong CH, Lo AW. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non–Small-Cell Lung Cancer. *JCO Clinical Cancer Informatics*. 2019;(3):1-11. doi:https://doi.org/10.1200/cci.19.00046

6.Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. *Journal of Nuclear Medicine*. 2020;61(4):488-495. doi:https://doi.org/10.2967/jnumed.118.222893

7.Sun W, Jiang M, Dang J, Chang P, Yin FF. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. *Radiation Oncology*. 2018;13(1). doi:https://doi.org/10.1186/s13014-018-1140-9

8.Liu C, Gong J, Yu H, Liu Q, Wang S, Wang J. A CT-Based Radiomics Approach to Predict Nivolumab Response in Advanced Non-Small-Cell Lung Cancer. *Frontiers in Oncology*. 2021;11. doi:https://doi.org/10.3389/fonc.2021.544339

9.None Shaila S. G, Sunanda Rajkumari, Vadivel Ayyasamy. Introducing the Deep Learning for Digital Age. *Advances in computational intelligence and robotics book series*. Published online July 24, 2019:317-333. doi:https://doi.org/10.4018/978-1-5225-9902-9.ch017

10.Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clinical Cancer Research*. 2019;25(11):3266-3275. doi:https://doi.org/10.1158/1078-0432.ccr-18-2495

11.Lai YH, Chen WN, Hsu TC, Lin C, Tsao Y, Wu S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports*. 2020;10:4679. doi:https://doi.org/10.1038/s41598-020-61588-w

12.Hou KY, Chen JR, Wang YC, et al. Radiomics-Based Deep Learning Prediction of Overall Survival in Non-Small-Cell Lung Cancer Using Contrast-Enhanced Computed Tomography. *Cancers (Basel)* . 2022;14(15):3798-3798. doi:https://doi.org/10.3390/cancers14153798

13.A. Elkrief, Phan K, L. Di Jorio, et al. 1394P Deep learning model to predict clinical outcomes in patients with advanced non-small cell lung cancer treated with immune checkpoint inhibitors. *Annals of Oncology*. 2020;31:S885-S885. doi:https://doi.org/10.1016/j.annonc.2020.08.1708

14.Lu CF, Liao CY, Chao HS, et al. A radiomics-based deep learning approach to predict progression free-survival after tyrosine kinase inhibitor therapy in non-small cell lung cancer. *Cancer Imaging*. 2023;23(1). doi:https://doi.org/10.1186/s40644-023-00522-5

15.Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Abstract 868: Prediction of clinical benefit to checkpoint blockade in advanced NSCLC patients using radiomics of PET/CT images. *Cancer Research*. 2020;80(16_Supplement):868-868. doi:https://doi.org/10.1158/1538-7445.am2020-868

16.Bareschino MA, Schettino C, Rossi A, et al. Treatment of advanced non small cell lung cancer. *Journal of thoracic disease*. 2011;3(2):122-133. doi:https://doi.org/10.3978/j.issn.2072-1439.2010.12.08

17.Corral J, Borras JM, Lievens Y. Utilisation of radiotherapy in lung cancer: A scoping narrative literature review with a focus on the introduction of evidence-based therapeutic approaches in Europe. *Clinical and Translational Radiation Oncology*. 2024;45:100717-100717. doi:https://doi.org/10.1016/j.ctro.2023.100717

18.Hanna TP, Shafiq J, Delaney GP, Vinod SK, Thompson SR, Barton MB. The population benefit of evidence-based radiotherapy: 5-Year local control and overall survival benefits. *Radiotherapy and Oncology*. 2018;126(2):191-197. doi:https://doi.org/10.1016/j.radonc.2017.11.004

19.Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Medicine*. 2018;15(11). doi:https://doi.org/10.1371/journal.pmed.1002711

20.Afshar P, Mohammadi A, Tyrrell PN, et al. $$\text {DRTOP}$$: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer. *Scientific Reports*. 2020;10(1). doi:https://doi.org/10.1038/s41598-020-69106-8

21.Peng J, Zhang X, Hu Y, et al. Deep learning to estimate response of concurrent chemoradiotherapy in non-small-cell lung carcinoma. *Journal of Translational Medicine*. 2024;22(1). doi:https://doi.org/10.1186/s12967-024-05708-4

22.Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Scientific Reports*. 2017;7(1):46349. doi:https://doi.org/10.1038/srep46349

23.Caii W, Wu X, Guo K, Chen Y, Shi Y, Chen J. Integration of deep learning and habitat radiomics for predicting the response to immunotherapy in NSCLC patients. *Cancer Immunology Immunotherapy*. 2024;73(8). doi:https://doi.org/10.1007/s00262-024-03724-3

24.Jia S, Jiang S, Lin Z, Li N, Xu M, Yu S. A survey: Deep learning for hyperspectral image classification with few labeled samples. *Neurocomputing*. 2021;448:179-204. doi:https://doi.org/10.1016/j.neucom.2021.03.035

25.Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*. 2021;15(1). doi:https://doi.org/10.1007/s12065-020-00540-3

26.Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. *Frontiers in Oncology*. 2020;10. doi:https://doi.org/10.3389/fonc.2020.580919

27.Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*. 2017;19(1):221-248. doi:https://doi.org/10.1146/annurev-bioeng-071516-044442

28.Wetzer E, Harlin H, Lindblad J, Natasa Sladoje. When Texture Matters: Texture-Focused Cnns Outperform General Data Augmentation and Pretraining in Oral Cancer Detection. Published online April 1, 2020. doi:https://doi.org/10.1109/isbi45749.2020.9098424

29.Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities. *IEEE Signal Processing Magazine*. 2019;36(4):132-160. doi:https://doi.org/10.1109/msp.2019.2900993

30. Kanakarajan, H., De Baene, W., Hanssens, P., & Sitskoorn, M. (2024). Predicting local control of brain metastases after stereotactic radiotherapy with clinical, radiomics and deep learning features. *Radiation Oncology*, *19*(1), 182.

31.Hou R, Xia W, Fu XL. Integrating CT Images and Radiotherapy Dose Distribution Maps to Predict the Progression-Free Survival in Locally Advanced Non-Small Cell Lung Cancer Patients. *International Journal of Radiation Oncology*Biology*Physics*. 2024;120(2):e26. doi:https://doi.org/10.1016/j.ijrobp.2024.07.1834

32. Zhao Y, Townsend A, Udell M. Probabilistic Missing Value Imputation for Mixed Categorical and Ordered Data. *arXiv (Cornell University)*. Published online January 1, 2022. doi:https://doi.org/10.48550/arxiv.2210.06673

33. Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning. Published online January 5, 2023. doi:https://doi.org/10.1007/s10994-022-06296-4

34.Hara K, Saito D, Shouno H. Analysis of function of rectified linear unit used in deep learning. 2015 International Joint Conference on Neural Networks (IJCNN). Published online July 2015. doi:https://doi.org/10.1109/ijcnn.2015.7280578

35. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*. 2012;60(6):84-90. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

36. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1). doi:https://doi.org/10.1186/s12864-019-6413-7

37. Braghetto A, Marturano F, Paiusco M, Baiesi M, Bettinelli A. Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset. Scientific Reports. 2022;12(1). doi:https://doi.org/10.1038/s41598-022-18085-z38.

38. Liao CY, Chen YM, Wu YT, et al. Personalized prediction of immunotherapy response in lung cancer patients using advanced radiomics and deep learning. Cancer Imaging. 2024;24(1). doi:https://doi.org/10.1186/s40644-024-00779-4

39. Kawaguchi T, Takada M, Kubo A, et al. Performance Status and Smoking Status Are Independent Favorable Prognostic Factors for Survival in Non-small Cell Lung Cancer: A Comprehensive Analysis of 26,957 Patients with NSCLC. Journal of Thoracic Oncology. 2010;5(5):620-630. doi:https://doi.org/10.1097/JTO.0b013e3181d2dcd9

40. Narra LR, Kumar R, Deek MP, Jabbour SK. Updates in Management of Unresectable Stage III Non Small Cell Lung Cancer: A Radiation Oncology Perspective. *Cancers (Basel)*. 2024;16(24):4233. doi:https://doi.org/10.3390/cancers16244233

41. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online June 2016:770-778. doi:https://doi.org/10.1109/cvpr.2016.90

42.Cheng X. A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. *Insights in computer, signals and systems*. 2024;1(1):65-78. doi:https://doi.org/10.70088/xpf2b276

43. Imani M, Beikmohammadi A, Arabnia HR. Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*. 2025;13(3):88. doi:https://doi.org/10.3390/technologies13030088

44.Fatima S, Hussain A, Sohaib Bin Amir, Syed Haseeb Ahmed, Syed. XGBoost and Random Forest Algorithms: An in Depth Analysis. *Pakistan journal of scientific research*. 2023;3(1):26-31. doi:https://doi.org/10.57041/pjosr.v3i1.946

45.Maeda R, Yoshida J, Ishii G, et al. The prognostic impact of cigarette smoking on patients with NSCLC. J Thorac Cardiovasc Surg. 2011 Apr;6(4):735-42. doi: 10.1097/JTO.0b013e318208e963.

46.Lee JH, Wu HG, Kim HJ, et al. Influence of comorbidities on the efficacy of radiotherapy with or without chemotherapy in elderly stage III NSCLC patients. Cancer Res Treat. 2012;44(4):242–250. doi:10.4143/crt.2012.44.4.242.

47.Antonia SJ, Villegas A, Daniel D, et al. Durvalumab after chemoradiotherapy in stage III NSCLC. N Engl J Med. 2017;377(20):1919–1929.doi: 10.1056/NEJMoa1709937

48.Ostheimer C, Eveslage M, Münch S, et al. Prognostic impact of gross tumor volume during radical radiochemotherapy of locally advanced NSCLC—results from the NCT03055715 multicenter cohort study of the Young DEGRO Trial Group. Strahlenther Onkol. 2021;197(5):385–395. doi:10.1007/s00066-020-01727-4.

49. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56

50. Caruso CM, Valerio Guarrasi, Ramella S, Soda P. A deep learning approach for overall survival prediction in lung cancer with missing values. Computer Methods and Programs in Biomedicine. 2024;254:108308-108308. doi:https://doi.org/10.1016/j.cmpb.2024.108308

51.Elbashir MK, Almotilag A, Mahmood MA, Mohammed M. Enhancing Non-Small Cell Lung Cancer Survival Prediction through Multi-Omics Integration Using Graph Attention Network. Diagnostics. 2024;14(19):2178. doi:https://doi.org/10.3390/diagnostics14192178

52. Astley JR, Reilly JM, Robinson S, Wild JM, Hatton MQ, Tahir BA. Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. *Radiotherapy and Oncology*. 2024;193:110084-110084. doi:https://doi.org/10.1016/j.radonc.2024.110084

53. Tang FH, Fong YW, Yung SH, Wong CK, Tu CL, Chan MT. Radiomics-Clinical AI Model with Probability Weighted Strategy for Prognosis Prediction in Non-Small Cell Lung Cancer. *Biomedicines*. 2023;11(8):2093-2093. doi:https://doi.org/10.3390/biomedicines11082093

54. Chang R, Qi S, Wu Y, Yue Y, Zhang X, Qian W. Nomograms integrating CT radiomic and deep learning signatures to predict overall survival and progression-free survival in NSCLC patients treated with chemotherapy. *Cancer Imaging*. 2023;23(1). doi:https://doi.org/10.1186/s40644-023-00620-4

55. Junaid KP, Kiran T, Gupta M, Kishore K, Sujata Siwatch. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Population Health Metrics*. 2025;23(1). doi:https://doi.org/10.1186/s12963-025-00364-2

56. Anand V, Gupta S, Gupta D, et al. Weighted Average Ensemble Deep Learning Model for Stratification of Brain Tumor in MRI Images. Diagnostics. 2023;13(7):1320-1320. doi:https://doi.org/10.3390/diagnostics13071320

57. Jabbar A, Serkan Savaş. A Weighted Ensemble Approach with Multiple Pre-Trained Deep Learning Models for Classification of Stroke. Medinformatics. Published online December 27, 2023. doi:https://doi.org/10.47852/bonviewmedin32021963

58. Lim E, Baldwin D, Beckles M, et al. Guidelines on the radical management of patients with lung cancer. Thorax. 2010;65(Suppl 3):iii1-iii27. doi:https://doi.org/10.1136/thx.2010.145938

59. Baldwin DR, White B, Schmidt-Hansen M, Champion AR, Melder AM. Diagnosis and treatment of lung cancer: summary of updated NICE guidance. BMJ. 2011;342. doi:https://doi.org/10.1136/bmj.d2110

60.Conibear J. Rationale for concurrent chemoradiotherapy for patients with stage III non-small-cell lung cancer. British Journal of Cancer. 2020;123(S1):10-17. doi:https://doi.org/10.1038/s41416-020-01070-6

61. West H (Jack), Jin JO. Performance Status in Patients With Cancer. *JAMA Oncology*. 2015;1(7):998. doi:https://doi.org/10.1001/jamaoncol.2015.3113

Appendix

Full list of Radiomics features:

original_shape_Elongation
original_shape_Flatness
original_shape_LeastAxisLength
original_shape_MajorAxisLength
original_shape_Maximum2DDiameterColumn
original_shape_Maximum2DDiameterRow
original_shape_Maximum2DDiameterSlice
original_shape_Maximum3DDiameter
original_shape_MeshVolume
original_shape_MinorAxisLength
original_shape_Sphericity
original_shape_SurfaceArea
original_shape_SurfaceVolumeRatio
original_shape_VoxelVolume
original_firstorder_10Percentile
original_firstorder_90Percentile
original_firstorder_Energy
original_firstorder_Entropy
original_firstorder_InterquartileRange
original_firstorder_Kurtosis
original_firstorder_Maximum
original_firstorder_MeanAbsoluteDeviation
original_firstorder_Mean
original_firstorder_Median
original_firstorder_Minimum
original_firstorder_Range
original_firstorder_RobustMeanAbsoluteDeviation
original_firstorder_RootMeanSquared
original_firstorder_Skewness
original_firstorder_TotalEnergy
original_firstorder_Uniformity
original_firstorder_Variance
original_glcm_Autocorrelation
original_glcm_ClusterProminence
original_glcm_ClusterShade
original_glcm_ClusterTendency
original_glcm_Contrast
original_glcm_Correlation
original_glcm_DifferenceAverage

original_glcm_DifferenceEntropy
original_glcm_DifferenceVariance
original_glcm_Id
original_glcm_Idm
original_glcm_Idmn
original_glcm_Idn
original_glcm_Imc1
original_glcm_Imc2
original_glcm_InverseVariance
original_glcm_JointAverage
original_glcm_JointEnergy
original_glcm_JointEntropy
original_glcm_MCC
original_glcm_MaximumProbability
original_glcm_SumAverage
original_glcm_SumEntropy
original_glcm_SumSquares
original_gldm_DependenceEntropy
original_gldm_DependenceNonUniformity
original_gldm_DependenceNonUniformityNormalized
original_gldm_DependenceVariance
original_gldm_GrayLevelNonUniformity
original_gldm_GrayLevelVariance
original_gldm_HighGrayLevelEmphasis
original_gldm_LargeDependenceEmphasis
original_gldm_LargeDependenceHighGrayLevelEmphasis
original_gldm_LargeDependenceLowGrayLevelEmphasis
original_gldm_LowGrayLevelEmphasis
original_gldm_SmallDependenceEmphasis
original_gldm_SmallDependenceHighGrayLevelEmphasis
original_gldm_SmallDependenceLowGrayLevelEmphasis
original_glrlm_GrayLevelNonUniformity
original_glrlm_GrayLevelNonUniformityNormalized
original_glrlm_GrayLevelVariance
original_glrlm_HighGrayLevelRunEmphasis
original_glrlm_LongRunEmphasis
original_glrlm_LongRunHighGrayLevelEmphasis
original_glrlm_LongRunLowGrayLevelEmphasis
original_glrlm_LowGrayLevelRunEmphasis
original_glrlm_RunEntropy
original_glrlm_RunLengthNonUniformity
original_glrlm_RunLengthNonUniformityNormalized
original_glrlm_RunPercentage
original_glrlm_RunVariance
original_glrlm_ShortRunEmphasis
original_glrlm_ShortRunHighGrayLevelEmphasis
original_glrlm_ShortRunLowGrayLevelEmphasis
original_glszm_GrayLevelNonUniformity
original_glszm_GrayLevelNonUniformityNormalized

original_glszm_GrayLevelVariance
original_glszm_HighGrayLevelZoneEmphasis
original_glszm_LargeAreaEmphasis
original_glszm_LargeAreaHighGrayLevelEmphasis
original_glszm_LargeAreaLowGrayLevelEmphasis
original_glszm_LowGrayLevelZoneEmphasis
original_glszm_SizeZoneNonUniformity
original_glszm_SizeZoneNonUniformityNormalized
original_glszm_SmallAreaEmphasis
original_glszm_SmallAreaHighGrayLevelEmphasis
original_glszm_SmallAreaLowGrayLevelEmphasis
original_glszm_ZoneEntropy
original_glszm_ZonePercentage
original_glszm_ZoneVariance
original_ngtdm_Busyness
original_ngtdm_Coarseness
original_ngtdm_Complexity
original_ngtdm_Contrast
original_ngtdm_Strength

Performance of XBG classifier for test data set:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC	MCC	Specificity (%)
Clinical features only	77.55	44.44	40.00	42.11	0.64	0.28	87.18
Clinical and deep learning features	73.47	33.33	30.00	31.58	0.57	0.15	84.62
Clinical and radiomics features	77.55	44.44	40.00	42.11	0.64	0.28	87.18
Clinical, radiomics and deep learning features	83.67	58.33	70.00	63.64	0.79	0.54	87.18

Performance of NN classifier for test data set:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC	MCC	Specificity (%)
Clinical features only	22.45	20.83	100.00	34.48	0.51	0.07	2.56
Clinical and deep learning features	24.49	21.28	100.00	35.09	0.53	0.10	5.13
Clinical and radiomics features	20.41	20.41	100.00	33.90	0.50	0.00	0.00
Clinical, radiomics and deep learning features	18.37	17.39	80.00	28.57	0.41	-0.29	2.56

Outcome of statistical analysis for Precision:

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	0.5897	0.6885
clinical features only vs clinical and radiomics features	0.0583	0.1522
clinical features only vs combination of all 3 features	0.1392	0.2088
clinical and deep learning features vs clinical and radiomics features	0.0761	0.1522
clinical and deep learning features vs combination of all 3 features	0.6885	0.6885
clinical and radiomics features vs combination of all 3 features	**0.0005**	**0.0028**

Outcome of statistical analysis for Recall:

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	**0.0084**	**0.0252**
clinical features only vs clinical and radiomics features	0.9195	0.9642
clinical features only vs combination of all 3 features	**0.0018**	**0.0108**
clinical and deep learning features vs clinical and radiomics features	0.1013	0.1520
clinical and deep learning features vs combination of all 3 features	0.9642	0.9642
clinical and radiomics features vs combination of all 3 features	0.0379	0.0757

Outcome of statistical analysis for F1 score:

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	0.1585	0.2378
clinical features only vs clinical and radiomics features	0.3750	0.4500
clinical features only vs combination of all 3 features	**0.0014**	**0.0043**
clinical and deep learning features vs clinical and radiomics features	0.0637	0.1273
clinical and deep learning features vs combination of all 3 features	0.6327	0.6327
clinical and radiomics features vs combination of all 3 features	**0.0002**	**0.0013**

Outcome of statistical analysis for AUC:

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	0.1585	0.2378
clinical features only vs clinical and radiomics features	0.3536	0.4244
clinical features only vs combination of all 3 features	**0.0014**	**0.0042**
clinical and deep learning features vs clinical and radiomics features	0.0581	0.1162
clinical and deep learning features vs combination of all 3 features	0.7024	0.7024
clinical and radiomics features vs combination of all 3 features	**0.0001**	**0.0008**

Outcome of statistical analysis for MCC:

Models	p-value	corrected p-value
clinical features only vs clinical and deep learning features	0.1917	0.2875
clinical features only vs clinical and radiomics features	0.3536	0.4244
clinical features only vs combination of all 3 features	**0.0014**	**0.0043**
clinical and deep learning features vs clinical and radiomics features	0.0581	0.1162
clinical and deep learning features vs combination of all 3 features	0.6327	0.6327
clinical and radiomics features vs combination of all 3 features	**0.0001**	**0.0008**

Top 10 features:

The top 10 features of the clinical only model are WHOPS, NSCLC stage, laterality, N classification, presence of cerebra vascular issues, durvalumab administered indicator, CRP (C-reactive protein) level, creatinine level, presence of cardiac surgery and presence of vascular issues.

The top 10 features of clinical with deep learning model are WHOPS, DL_426, DL_98, DL_202, DL_288, DL_384, DL_30, DL_414, DL_301 and DL_122.

The top 10 features of clinical with radiomics model are WHOPS, NSCLC stage, presence of cardiac surgery, presence of cerebra vascular issues, creatinine level, N classification, CRP (C-reactive protein) level, laterality, radiotherapy treatment dose, and original_shape_MinorAxisLength.

ROC Curves:

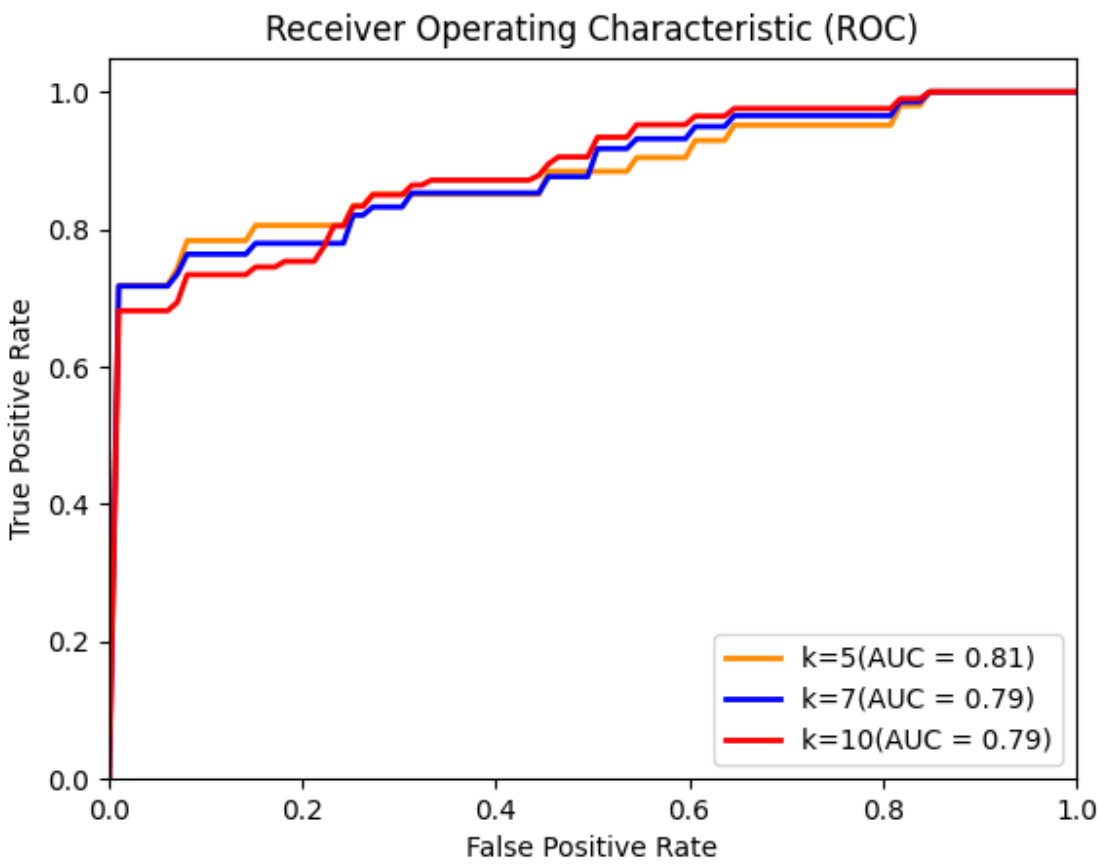

Figure 4: ROC curve of model trained with clinical features only.

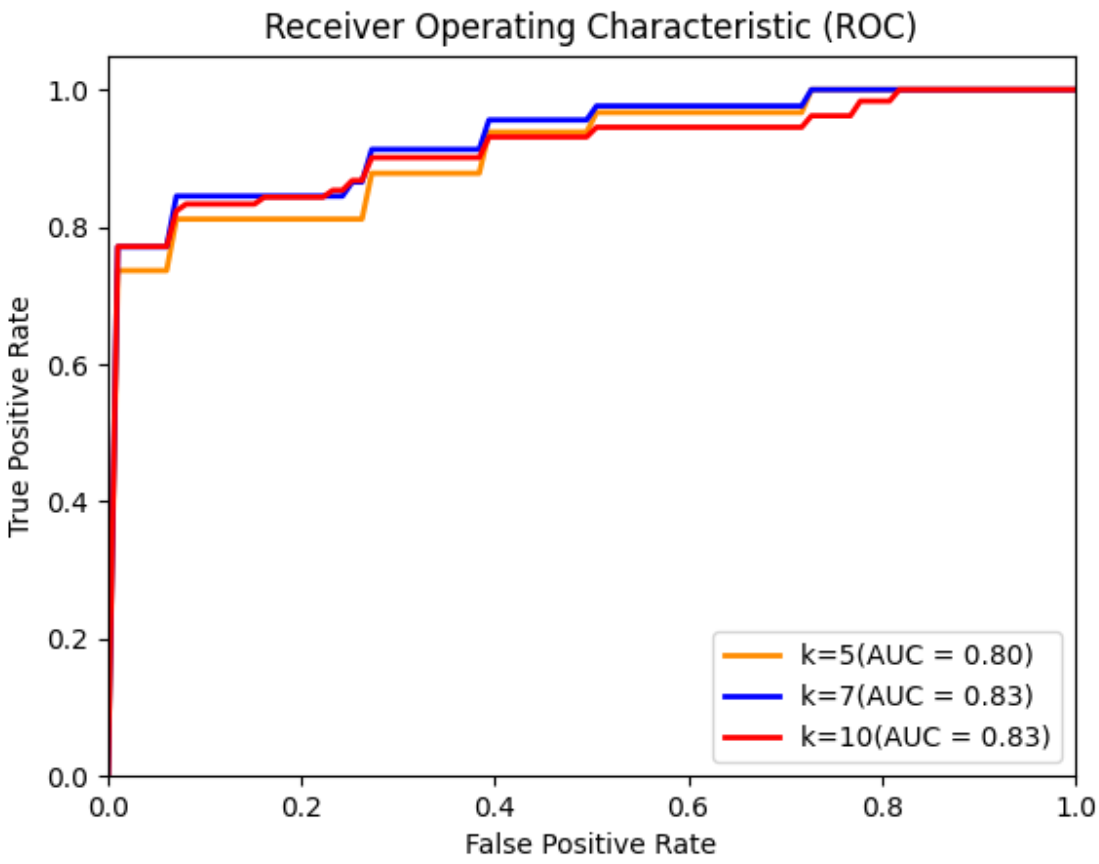

Figure 5: ROC curve of model trained with clinical and deep learning features.

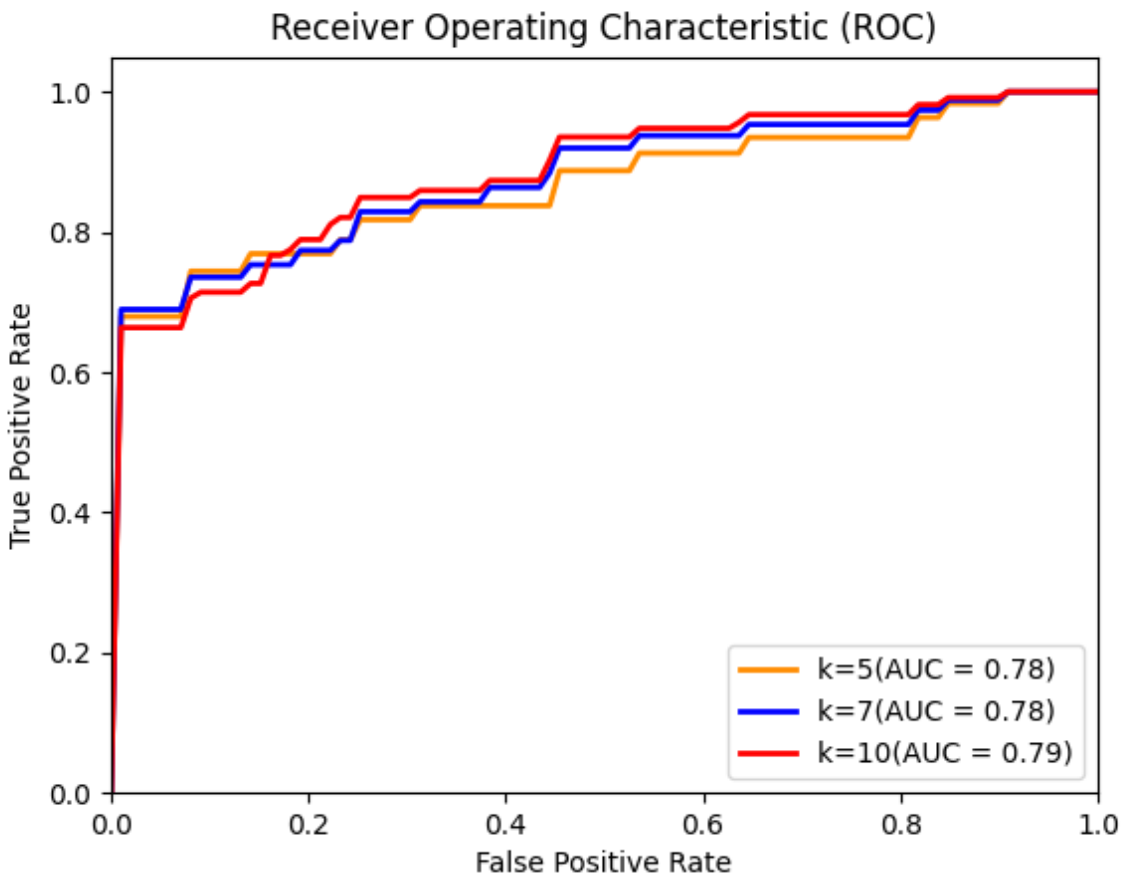

Figure 6: ROC curve of model trained with clinical and radiomics features.