Physical and Engineering Sciences in Medicine
https://doi.org/10.1007/s13246-025-01650-x

SCIENTIFIC PAPER

Machine learning–assisted classification of lung cancer: the role of sarcopenia, inflammatory biomarkers, and PET/CT anatomical-metabolic parameters

Handan Tanyildizi-Kokkulunk[1] · Goksel Alcin[2] · Iffet Cavdar[3] · Resit Akyel[4] · Safak Yigit[5] · Tuba Ciftci-Kusbeci[6] · Gonul Caliskan[7]

Received: 20 May 2025 / Accepted: 10 September 2025

Abstract

Accurate differentiation between non-cancerous, benign, and malignant lung cancer remains a diagnostic challenge due to overlapping clinical and imaging characteristics. This study proposes a multimodal machine learning (ML) framework integrating positron emission tomography/computed tomography (PET/CT) anatomic-metabolic parameters, sarcopenia markers, and inflammatory biomarkers to enhance classification performance in lung cancer. A retrospective dataset of 222 patients was analyzed, including demographic variables, functional and morphometric sarcopenia indices, hematological inflammation markers, and PET/CT derived parameters such as maximum and mean standardized uptake value (SUVmax, SUVmean), metabolic tumor volume (MTV), total lesion glycolysis (TLG). Five ML algorithms—Logistic Regression, Multi-Layer Perceptron, Support Vector Machine, Extreme Gradient Boosting, and Random Forest—were evaluated using standardized performance metrics. Synthetic Minority Oversampling Technique was applied to balance class distributions. Feature importance analysis was conducted using the optimal model, and classification was repeated using the top 15 features. Among the models, Random Forest demonstrated superior predictive performance with a test accuracy of 96%, precision, recall, and F1-score of 0.96, and an average AUC of 0.99. Feature importance analysis revealed SUVmax, SUVmean, total lesion glycolysis, and skeletal muscle index as leading predictors. A secondary classification using only the top 15 features yielded even higher test accuracy (97%). These findings underscore the potential of integrating metabolic imaging, physical function, and biochemical inflammation markers in a non-invasive ML-based diagnostic pipeline. The proposed framework demonstrates high accuracy and generalizability and may serve as an effective clinical decision support tool in early lung cancer diagnosis and risk stratification.

Keywords Lung cancer classification · Machine learning · Random forest · PET/CT · Sarcopenia · Inflammatory biomarkers

Abbreviations

Demographic / Anthropometric variables

H	Height (m)
W	Weight (kg)
BMI	Body mass index (kg/m²)
NS	Never smokers
FS	Former smokers
CS	Current smokers

Sarcopenia-related parameters

F-sarcopenia	Functional sarcopenia
M-sarcopenia	Morphological sarcopenia
SMA	Skeletal muscle area (cm²)
SMI	Skeletal muscle index (cm²/m²)
TUG	Timed up and go (s)
FTSST	Five times sit-to-stand test (s)

Inflammatory / Biochemical markers

ALB	Albumin (g/L)

Extended author information available on the last page of the article

Published online: 06 October 2025

BIL	Bilirubin (mg/dL)
CRP	C-reactive protein (mg/L)
AST	Aspartate aminotransferase (U/L)
ALT	Alanine aminotransferase (U/L)
NEU	Neutrophil ($10^3/\mu L$)
MON	Monocyte ($10^3/\mu L$)
WBC	White blood cell count ($10^3/\mu L$)
ALB/BIL	Albumin-to-Bilirubin ratio
ALB/CRP	Albumin-to-CRP ratio
ALB/AST	Albumin-to-AST ratio
ALB/ALT	Albumin-to-ALT ratio
CRP/BIL	CRP-to-Bilirubin ratio
AST/BIL	AST-to-Bilirubin ratio
ALT/BIL	ALT-to-Bilirubin ratio
CRP/AST	CRP-to-AST ratio
AST/ALTT	AST-to-ALT ratio
NEU/MON	Neutrophil-to-monocyte ratio
NEU/WBC	Neutrophil-to-WBC ratio
MON/WBC	Monocyte-to-WBC ratio
CRP/WBC	CRP-to-WBC ratio

Positron emission tomography/computed (PE/CT) parameters

ROI	Region of Interest
SUVmean	Mean standardized uptake value
SUVmax	Maximum standardized uptake value
SUVort	Average lesion SUV
MTV	Metabolic tumor volume
TLG	Total lesion glycolysis

Machine learning and statistics

ML	Machine learning
LR	Logistic regression
SVM	Support vector machine
MLP	Multi-layer perceptron
RF	Random forest
XGBoost	Extreme gradient boosting
SMOTE	Synthetic minority oversampling technique
AUC	Area under the curve

Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide, accounting for approximately 1.8 million deaths annually [1]. Accurately distinguishing benign from malignant pulmonary lesions is essential for guiding treatment but remains challenging due to overlapping clinical, radiological, and metabolic features [2].

The gold standard for diagnosis is histopathological confirmation via tissue biopsy. While definitive, biopsies are invasive, costly, and sometimes infeasible due to patient comorbidities or inaccessible lesion locations. In this context, non-invasive diagnostic adjuncts are gaining importance. Among these, systemic inflammatory biomarkers—such as the neutrophil-to-lymphocyte ratio, monocyte-to-lymphocyte ratio, and systemic immune-inflammation index—have demonstrated associations with tumor progression, immune status, and prognosis in lung cancer [3–5].

Positron emission tomography/computed tomography (PET/CT) has become a key imaging modality for the evaluation of pulmonary lesions. Quantitative PET/CT derived parameters—including maximum standardized uptake value (SUVmax), metabolic tumor volume (MTV), and total lesion glycolysis (TLG)—provide valuable insights into tumor metabolism and burden. Numerous studies have reported their prognostic and diagnostic value in lung cancer staging and treatment response [6–8].

On the other hand, sarcopenia, characterized by progressive loss of skeletal muscle mass and function, has emerged as an important prognostic factor in oncology [9]. In lung cancer, sarcopenia is linked to increased treatment toxicity, reduced survival, and greater risk of postoperative complications [10]. Both morphometric assessments (e.g., skeletal muscle index, SMI) and functional tests (e.g., gait speed, handgrip strength) can be derived from clinical data or imaging, providing a non-invasive measure of patient frailty and physical reserve [11].

Machine learning (ML) techniques have shown strong potential in improving diagnostic accuracy, risk stratification, and treatment planning in oncology [12–18]. Prior ML-based lung cancer classification studies have typically relied on a single data modality—such as imaging [19–22], blood-based biomarkers [23], or radiomics-based sarcopenia assessment [24]. While these approaches have achieved promising accuracy, they may overlook complementary information available from other domains.

The novelty of this study is the combination of three complementary and clinically significant data domains—PET/CT-based anatomical-metabolic indicators, systemic inflammatory markers, and both functional and morphometric measures of sarcopenia—within a unified multimodal ML framework for lung cancer classification. To the best of our knowledge, no prior work has combined these modalities to classify non-cancerous, benign, and malignant lesions, nor examined the extent to which such a multimodal dataset can improve diagnostic accuracy. This study addresses this gap by developing and evaluating a model that leverages the complementary strengths of these data sources to enhance diagnostic precision and support clinical decision-making in lung cancer management.

Literature review

Understanding the diagnostic landscape of lung cancer requires integrating insights from different research domains. Prior studies can broadly be grouped into three categories: (i) imaging based ML approaches, which extract radiological features from CT or PET/CT scans; (ii) blood derived inflammatory biomarkers, which provide systemic and easily accessible signals of disease; and (iii) sarcopenia related measures, reflecting patient frailty and metabolic status. Reviewing these domains separately allows us to highlight their distinct contributions and limitations, while also underscoring the need for a multimodal integration strategy.

Imaging based machine learning approaches

Imaging data, particularly from CT and PET/CT scans, remain the cornerstone of lung cancer diagnostics. Advanced ML and deep learning frameworks have shown strong potential in this domain. For example, Hendrix et al. [19] developed an AI system to detect lung nodules with high sensitivity, comparable to radiologists, using CT scans [19]. Similarly, Lv et al. [20] proposed MIFNet, a multi-scale deep learning model, reporting excellent sensitivity and specificity for CT-based classification [20]. Radiomic approaches, such as those by Delzell et al. [22], have further illustrated the feasibility of extracting high-dimensional features for ML classifiers [22].

Despite these successes, imaging-only models tend to operate in isolation, focusing narrowly on radiological signals. While they excel at identifying lesion morphology and texture, they may miss systemic physiological or biochemical factors that also shape disease status. This limitation motivates the exploration of complementary modalities.

ML models using blood derived inflammatory biomarkers

Parallel to imaging studies, a growing body of work has investigated the role of blood-derived indices in cancer detection and prognosis. These biomarkers are attractive because they are cost-effective, minimally invasive, and already embedded in clinical workflows. [23] reported that AI models using routine blood indices achieved up to 95.7% accuracy in cancer detection [23]. In addition, systemic immune-inflammation indices have been repeatedly linked to treatment response and survival outcomes in lung cancer [3, 5]. However, while blood-based approaches offer systemic insight, they lack spatial and anatomical context. A high inflammatory index may indicate cancer-related processes but cannot localize or characterize lesions. This trade-off underscores the complementary value of imaging data, suggesting that integration across modalities may yield a more complete diagnostic picture.

Applications of sarcopenia related measures in oncology

Sarcopenia, assessed through morphometric (M-sarcopenia) and functional (F-sarcopenia) measures, has been shown to impact cancer prognosis. Yang et al. [10] identified sarcopenia as an independent predictor of poor outcomes in lung cancer [10]. More recently, Dong et al. [24] utilized chest CT radiomics to detect sarcopenia in advanced non-small cell type lung cancer patients with a LightGBM model, achieving 90% accuracy [24]. Yet, sarcopenia remains underexplored in diagnostic classification studies. Most prior work has emphasized its prognostic implications, while its diagnostic potential—particularly when combined with imaging features and systemic biomarkers—has not been fully realized. Thus, sarcopenia represents a promising but underutilized dimension in multimodal ML frameworks.

Towards a multimodal framework

Taken together, existing studies demonstrate impressive advances within their individual domains. Imaging-based ML models excel at morphological characterization; blood-derived biomarkers provide systemic, accessible indicators; and sarcopenia measures add prognostic and physiological context. However, the literature largely treats these modalities in isolation, with limited attempts at systematic integration [12–14].

This fragmentation reveals a clear research gap: no study to date has developed a unified diagnostic framework that incorporates PET/CT derived anatomical-metabolic parameters, systemic inflammatory biomarkers, and sarcopenia measures to classify non-cancerous, benign, and malignant lesions. Our study addresses this gap by proposing a multimodal ML framework that synthesizes these complementary data sources. By combining diverse but clinically relevant predictors, we aim to demonstrate that a carefully selected, reduced feature set can maintain high diagnostic accuracy while enhancing interpretability and clinical applicability. Table 1 presents a summary of the relevant literature.

Methods

This study was conducted in accordance with the ethical standards of the Altınbaş University Ethics Committee (Protocol No: 2022/157) and the principles outlined in the Declaration of Helsinki. Informed consent was obtained from all participants prior to their inclusion in the study. Figure 1

Table 1 The summary of representative studies across the three domains, illustrating both their strengths and their limitations when considered separately

Study	Year	Data modality	Method/ Model	Reported performance	Limitations
Hendrix et al. [19]	2023	CT imaging	AI system for nodule detection	High sensitivity, comparable to radiologists	Focused solely on imaging; no systemic or prognostic factors considered
Lv et al. [20]	2024	CT imaging	MIFNet (deep learning)	94.82% accuracy, 97.34% F1-score	Excellent CT performance, but lacks integration with biochemical or clinical data
Delzell et al. [22]	2019	CT radiomics	Elastic net, SVM	AUC 0.72	Moderate accuracy; radiomics used in isolation without systemic context
Kotoulas et al. [23]	2025	Blood indices	AI models	95.7% accuracy	Lacks anatomical specificity; cannot localize or characterize lesions
Guo et al. [3]	2018	Blood biomarkers	Prognostic analysis	Significant survival associations	Prognostic focus; not designed for diagnostic classification
Su et al. [5]	2025	Blood biomarkers	Meta-analysis	Predictive utility in treatment response	Aggregated evidence, but no integration with imaging modalities
Yang et al. [10]	2019	Sarcopenia (CT morphometry)	Meta-analysis	Prognostic value confirmed	Highlights importance of sarcopenia, but not applied in diagnostic ML frameworks
Dong et al. [24]	2020	Chest CT radiomics (sarcopenia detection)	LightGBM	90% accuracy	Sarcopenia detected via CT, but study limited to advanced cases and not combined with biomarkers

presents an overview of the data collection procedure and the step-by-step methodology applied during the ML.

Descriptions of the dataset

The study included patients who underwent PET/CT (Discovery 600, GE Medical Systems, Milwaukee, Wis, USA) imaging and routine clinical assessment at Nuclear Medicine Department of Yedikule Chest Diseases Hospital, Istanbul, Türkiye. The data were collected between 2023 and 2025.

Inclusion criteria were: (i) confirmed diagnosis of lung lesion (benign or malignant) or absence of malignancy, (ii) availability of PET/CT imaging parameters, laboratory results, and sarcopenia assessments. Patients with incomplete records were excluded. The final dataset consisted of 222 patients (F: 58, M: 164), categorized into three groups: non-cancerous, benign, and malignant.

The dataset included various types of features grouped into four main categories:

(1) The demographic and anamnesis-based variables such as sex, age, height, weight, derived ratios as body mass index, and smoking status;

(2) The sarcopenia-related parameters derived from F-sarcopenia and M-sarcopenia assessments, including handgrip strength, quadriceps strength, gait speed, Timed Up and Go (TUG), Five Times Sit-to-Stand Test (FTSST), skeletal muscle area (SMA), and SMI;

(3) The next-generation inflammatory biomarkers obtained from blood tests, including albümin (ALB), bilirubin (BIL), C-reactive protein (CRP), aspartate aminotransferase (AST), etc. and.

(4) PET/CT derived anatomical-metabolic imaging features, including SUVmax, SUVmean, MTV, and TLG.

PET/CT metrics were obtained from manually delineated three-dimensional region of interest (ROIs) of the primary lesion on attenuation-corrected PET images, with CT used for anatomical verification. MTV was calculated using a fixed SUV threshold, and TLG as MTV×SUVmean. To verify volumetric reproducibility, two nuclear medicine physicians independently delineated a phantom of known volume (3025 cm^3). Measured volumes were 3045 cm^3 and 3037 cm^3, corresponding to absolute deviations of 20 cm^3 (0.66%) and 12 cm^3 (0.40%), with a mean deviation of 0.53%. These small errors support the reproducibility of volumetric measurements.

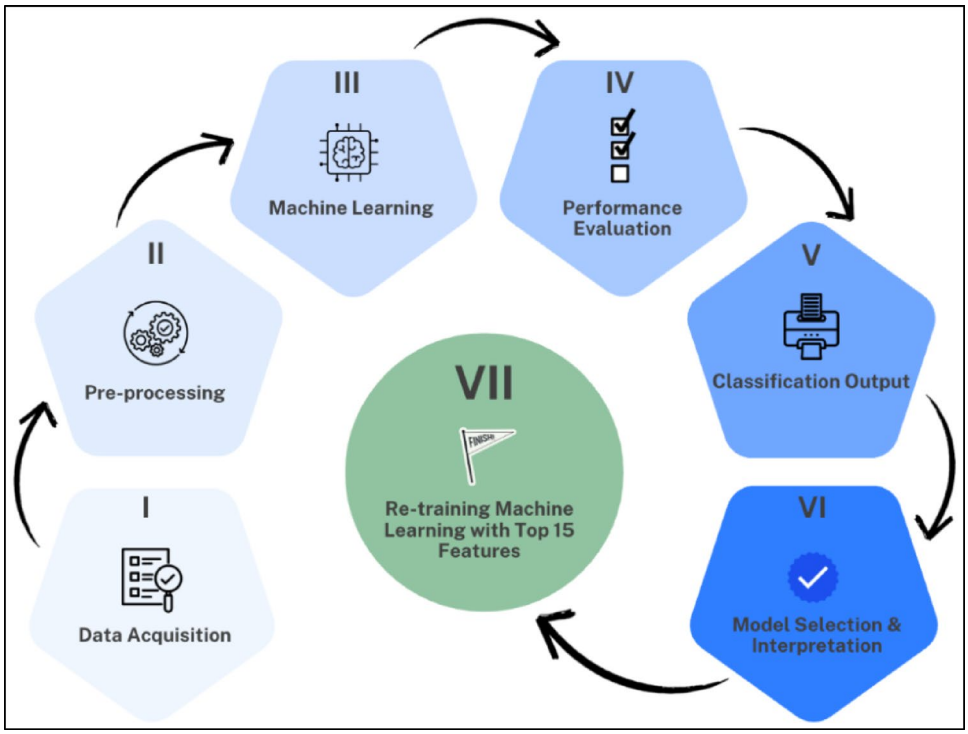

Fig. 1 The workflow of the study methodology which covers the entire process from data acquisition to ML-based classification and interpretation of lung cancer

The complete set of patient variables, together with the descriptive characteristics of the 222 patients and the class distributions (non-cancerous, benign, malignant), is presented in the list of abbreviations and the supplementary section.

Machine learning

Data preprocessing

To classify lung cancer into non-cancerous, benign, and malignant categories, a ML pipeline was implemented using Python (v3.9). The dataset was randomly split into training (70%) and testing (30%) subsets using the train_test_split function from Scikit-learn. The final performance metrics reported correspond exclusively to the independent test set, ensuring that no information from the test set was used during training or parameter optimization.

Prior to modeling, all numerical features were standardized using the StandardScaler (scikit-learn), which centers each feature by subtracting the mean and scales it to unit variance. Since the dataset contained no missing values, no imputation was necessary. Standardization parameters were computed from the training set and subsequently applied to the test set to avoid data leakage.

Given the class imbalance among diagnostic categories, the Synthetic Minority Oversampling Technique (SMOTE) was applied via the imblearn.over_sampling library. Class distribution in a representative outer-CV training split was imbalanced (non-cancerous = 44, benign = 44, malignant = 89). We applied SMOTE (k = 5, random_state = 2025) exclusively on training folds, balancing the classes to 89/89/89 and preventing test-set leakage. All randomized procedures used fixed seeds for reproducibility.

Model selection

The ML models employed in this study—Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), and Random Forest (RF)—were chosen to represent complementary learning paradigms encompassing linear models, kernel-based approaches, neural networks, gradient boosting, and bagging ensembles. Such diversity is particularly advantageous for heterogeneous, mixed-type clinical and imaging datasets, as it enables evaluation across methods with distinct inductive biases. LR and SVM have been extensively applied in biomedical classification tasks due to their interpretability and robustness in high-dimensional data spaces [25, 26]. LR estimates the probability of class membership using the logistic function, modeling the log odds of the outcome as a linear combination of input features [27]. SVM aims to find an optimal separating hyperplane that maximizes the margin between classes, with kernel functions enabling non-linear decision boundaries [28]. MLP provides the capacity to model complex, non-linear feature interactions, making it suitable for multimodal data integration [29, 30]. XGBoost is recognized for

Table 2 Hyperparameter ranges explored during grid search and final optimal values selected for each model

Model	Parameters & ranges explored	Final optimal values
RF	n_estimators = {100, 300, 500, 700} max_depth = {5, 10, 15}	n_estimators=500 max_depth=10
XGBoost	n_estimators = {100, 300, 500} max_depth = {3, 6, 10} learning_rate = {0.01, 0.05, 0.1} subsample = {0.6, 0.8, 1.0}	n_estimators=500 max_depth=6 learning_rate=0.1 subsample=0.8
SVM	C = {0.1, 1, 10} gamma = {0.001, 0.01, 0.1}	C=10 gamma=0.01
MLP	hidden_layer_sizes = {(50,), (100,), (100,50)} activation = {ReLU, tanh} learning_rate_init = {0.001, 0.01} alpha = {0.0001, 0.001}	hidden layers = (100, 50) activation=ReLU learning_rate_init=0.001 alpha=0.0001
LR	penalty = {L1, L2} C = {0.1, 1, 10} solver = {liblinear, saga}	penalty=L2 C=1.0 solver=liblinear

its efficiency and strong predictive performance in handling sparse and non-linear data [31], while RF offers robustness to multicollinearity and the capability to rank feature importance effectively [32]. This model spectrum facilitates a modality-agnostic and balanced performance comparison under identical preprocessing and validation settings.

Hyperparameter tuning

Hyperparameter tuning and model selection were performed exclusively on the training set using nested 5-fold cross-validation, and the final performance of each optimized model was then evaluated on the independent test set. This ensured that no information from the test set or validation folds leaked into the model training process, strengthening reproducibility and transparency.

Grid search was chosen because it provides exhaustive evaluation of predefined parameter ranges and, given the moderate dataset size and limited parameter space, offered a transparent and reproducible strategy compared to random optimization.

The parameter ranges explored and the final optimal values selected for each model are presented in Table 2. Model selection was based on both accuracy and AUC within the nested CV framework, with AUC prioritized when discrepancies occurred, given its threshold-independent evaluation of classification performance.

Model evaluation

Model performance was evaluated on the independent test set using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). In the multiclass setting, precision, recall (sensitivity), and F1-score were computed for each class using a one-vs-rest approach, and the macro-averaged values were reported. In addition, confusion matrices and ROC curves were generated to visually assess classification efficacy across classes.

Based on superior accuracy and AUC performance, the optimal classifier was identified. Subsequently, a feature importance analysis was conducted using this model, and the top 15 most predictive features were selected. The ML training process was then repeated using only these selected features to evaluate their discriminative power in reduced-dimension models.

All analyses and visualizations (e.g., ROC plots and confusion matrices) were performed using pandas, numpy, scikit-learn, matplotlib, seaborn, and imbalanced-learn libraries.

Table 3 Comparative performance metrics of ML algorithms for multiclass classification of lung cancer

	LR	MLP	SVM	XGBoost	RF
Train accuracy (%)	94.42	96	95.7	100	100
Test accuracy (%)	79	86	83	94	96
Precision	0.80	0.87	0.85	0.94	0.96
Recall	0.79	0.86	0.83	0.94	0.96
F1-score	0.78	0.86	0.83	0.94	0.96
AUC (average)	0.91	0.95	0.95	0.99	0.99
AUC (class 0)	0.95	1	0.98	1	1
AUC (class 1)	0.84	0.92	0.91	0.99	0.99
AUC (class 2)	0.95	0.96	0.98	0.99	0.99

Results

This section presents the outcomes of our ML analysis in three parts: (i) overall model performance across classifiers, (ii) statistical comparisons between models, and (iii) feature importance and reduced feature set classification.

Comparative performance of classifiers

The performance of five supervised ML algorithms—LR, MLP, SVM, XGBoost, and RF—was evaluated for multiclass classification of lung cancer. Metrics including training and test accuracy, precision, recall, F1-score, and AUC were used for model comparison, as summarized in Table 3. All performance results reported in Table 3 correspond to the independent 30% test set after hyperparameter optimization within the training set using nested 5-fold cross-validation. The nested CV ensured robust model selection, while the independent test evaluation provided an unbiased estimate of generalization performance.

LR, while yielding a reasonable training accuracy of 94.42%, exhibited a substantial decline in performance on the test set (79%), reflecting a clear overfitting issue. The model's discriminatory ability was also uneven across classes, with notably reduced performance for class 1 (AUC = 0.84), indicating difficulty in distinguishing benign group from other categories.

MLP showed initially promising results with perfect training accuracy (96%); however, the considerable drop to 86% in test accuracy (Δ = 10%) suggests significant overfitting and limited generalizability. For the MLP classifier, the training and validation curves are provided in Figs. 2 and 3. The training loss exhibits smooth convergence, while the validation accuracy stabilizes without oscillations, supporting the reliability of the reported MLP performance.

SVM delivered moderately balanced results, but its test accuracy of 83% remained lower than more advanced models. Although class-level AUCs were relatively high, the presence of overfitting (Δ = 12%) suggests that the model's ability to generalize was compromised, especially in distinguishing benign cases (AUC = 0.91).

XGBoost performed substantially better than the previous models, with perfect training accuracy (100%) and a test accuracy of 94% with minimal overfitting risk (Δ = 0.06). However, despite strong overall performance and minimal overfitting, the model did not surpass the final classifier in this study in terms of precision, recall, or test accuracy.

RF yielded the highest overall performance among all tested models, with a training accuracy of 100% and a test accuracy of 96%, and no signs of overfitting. The model achieved perfect precision (0.96) and the highest recall (0.96), resulting in an F1-score of 0.96. Its average AUC was 0.99, and class-specific AUCs reached 1.00 (non-cancerous), 0.99 (benign), and 0.99 (malignant). These results indicate that RF provided the most accurate and generalizable predictions in this multi-class classification task [33]. Table 4 presents class-wise performance metrics of the RF classifier on the independent test set.

Statistical comparison of models

To further examine whether differences across models were significant, a Friedman test was applied across the five classifiers, yielding significant overall differences ($\chi^2 = 15.6$, $p = 0.0013$). Post-hoc Nemenyi analysis confirmed that RF achieved significantly higher accuracy compared to other classifiers ($p < 0.05$). These statistical tests reinforce that RF was not only the top performer numerically but also statistically superior to alternative models.

Residual misclassifications mainly occurred between benign and malignant cases. For RF, confusion matrix totals showed non-cancerous cases were all correctly classified (55/55), while 7 benign cases were misclassified as malignant and 11 malignant cases were misclassified as benign.

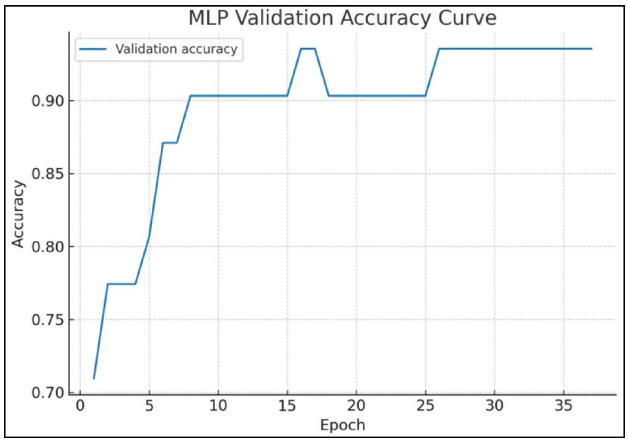

Fig. 2 Training and validation accuracy curves for the MLP classifier

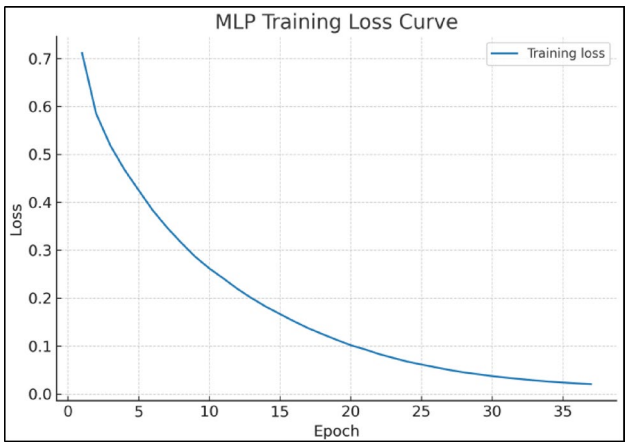

Fig. 3 Training and validation loss curves for the MLP classifier

Table 4 Class-wise performance metrics (precision, recall, F1-score, and AUC) of the best-performing RF classifier on the independent test set

Class	Precision	Recall	F1-score	AUC
Non-cancerous	1.00	1.00	1.00	1.00
Benign	0.96	0.88	0.94	0.99
Malignant	0.94	1.00	0.96	0.99

Feature importance analysis

Given its superior performance, RF was selected for feature importance analysis. The ranked importance values of the top predictive features are visualized in Fig. 4, providing insights into which variables played the most critical roles in differentiating between non-cancerous, benign, and malignant cases.

The feature importance analysis based on the RF model revealed that a small subset of variables had a disproportionately high impact on classification performance. The

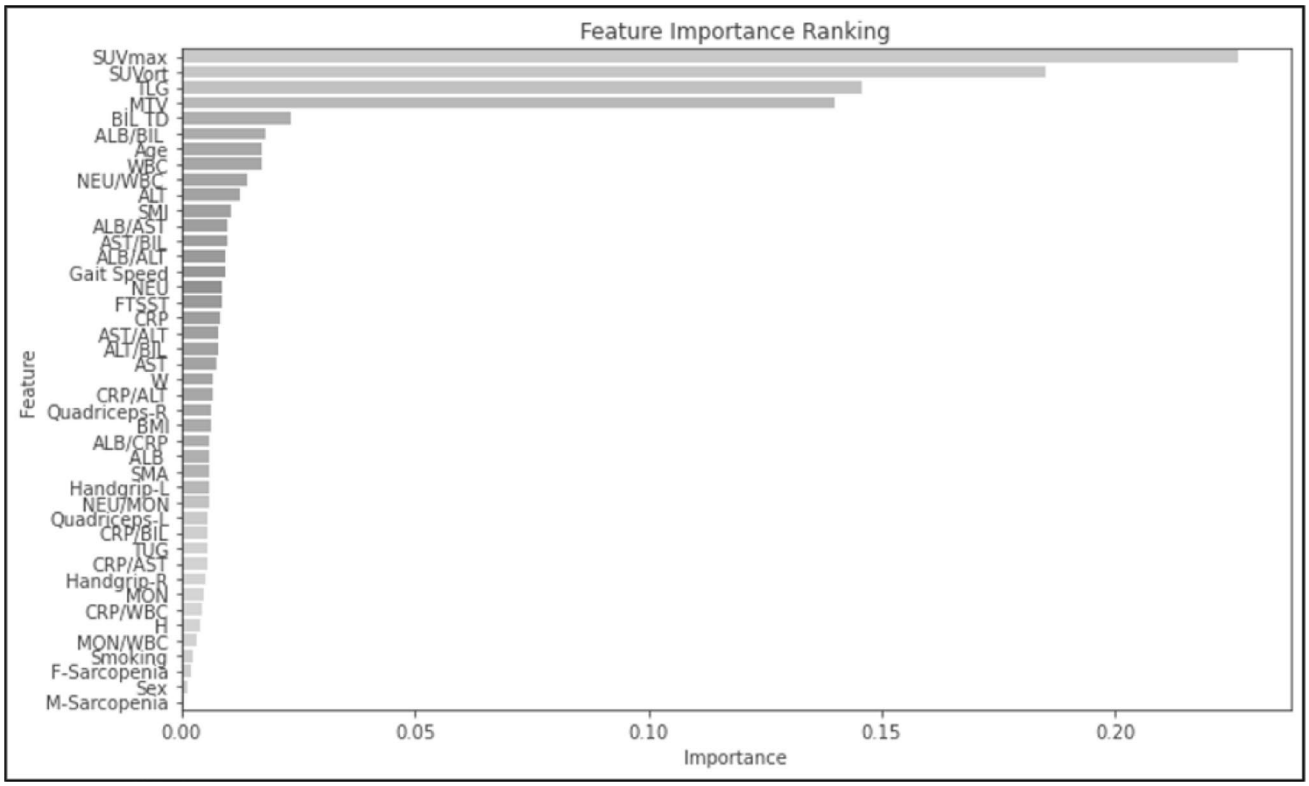

Fig. 4 Feature importance ranking based on the RF classifier

most influential feature was SUVmax, contributing 22.6% to the model's decisions, followed by SUVort (18.5%), TLG (14.6%), and MTV (14.0%). These PET/CT-derived parameters together accounted for nearly 69% of the model's predictive power, underscoring the central role of anatomical-metabolic imaging in lung cancer classification.

Additional features with moderate contribution included BIL (2.4%), ALB/BIL ratio (1.8%), age (1.7%), WBC (1.7%), NEU/WBC ratio (1.4%), and ALT (1.2%). Other features such as SMI (1.1%), ALB/AST (1.0%), AST/BIL (1.0%), ALB/ALT (0.9%), and gait speed (0.9%) also showed measurable influence on the classification output, indicating the added value of sarcopenia- and inflammation-related markers when used in conjunction with imaging variables.

To better understand the contribution of different feature groups, we explicitly aggregated RF–based feature importance values by category. PET/CT-derived variables accounted for the majority of predictive power with a mean contribution of approximately 69%, compared to 18% for inflammatory biomarkers and 13% for sarcopenia-related measures. This distribution highlights one of the key contributions of our study: although PET/CT parameters dominate classification performance, systemic inflammatory and sarcopenia measures provide complementary value, underscoring the importance of multimodal integration.

Classification using the top 15 features

To test whether performance could be preserved with fewer predictors, the RF model was retrained using only the top 15 features identified in the importance ranking. As shown in Fig. 5, the reduced model maintained high performance: training accuracy remained at 100%, while test accuracy increased slightly to 97%. Precision, recall, and F1-score all reached 0.97, with an average AUC of 0.99.

These findings indicate that a smaller, well-curated feature set can preserve near-optimal performance, improving interpretability and efficiency without sacrificing accuracy.

Statistical test results before and after feature selection

To assess the robustness of classifier performance, statistical tests were performed on both the full feature set and the reduced set of the top 15 most important features. For the full feature set, the Friedman test revealed significant overall differences among classifiers (χ^2=15.6, p=0.0013). Post-hoc Nemenyi analysis indicated that RF significantly outperformed all other models (p<0.05).

After feature selection, the Friedman test again showed significant differences (χ^2=16.2, p=0.0009), with RF maintaining a statistically significant advantage over LR, SVM,

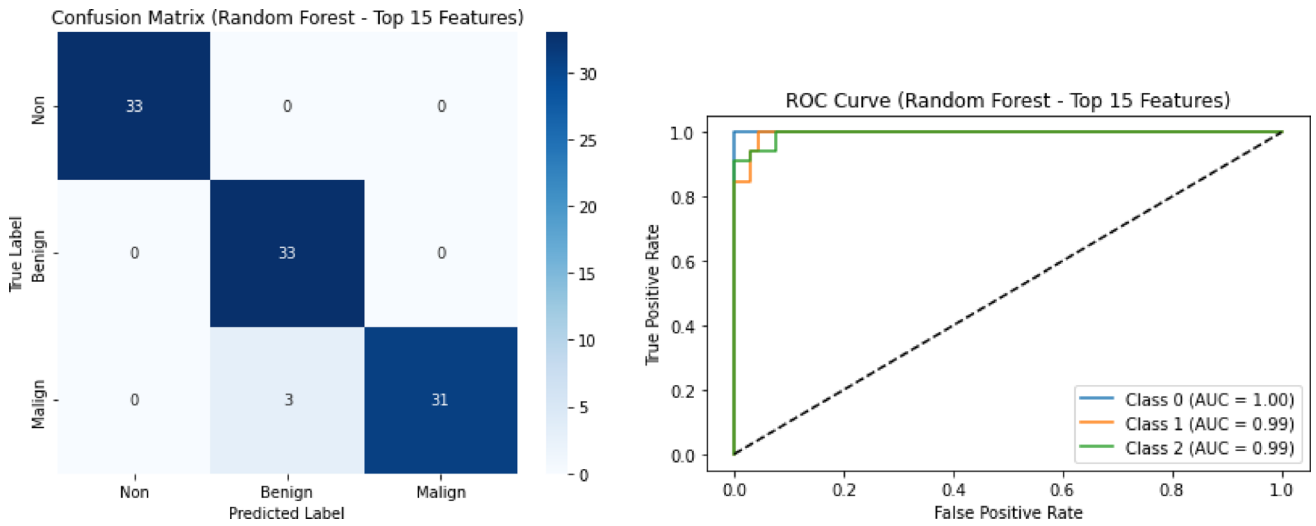

Fig. 5 Confusion matrix and ROC curves for RF classification using top 15 features

Table 5 Statistical test results comparing classifiers before and after feature selection

Condition	Test performed	χ^2 / Statistic	p-value	Post-hoc findings (Nemenyi)
Full feature set	Friedman	$\chi^2 = 15.6$	0.0013	RF significantly > LR, SVM, MLP, XGBoost
Reduced (Top 15 features)	Friedman	$\chi^2 = 16.2$	0.0009	RF significantly > LR, SVM, MLP, XGBoost

MLP, and XGBoost ($p < 0.05$). These results confirm that reducing the feature space did not diminish statistical separability but rather reinforced the stability and superiority of RF under both conditions. Table 5 summarizes the statistical outcomes.

Summary of results

The obtained results demonstrate that integrating PET/CT derived anatomical-metabolic parameters, systemic inflammatory biomarkers, and sarcopenia-related indices into a multimodal ML framework substantially improves classification performance compared to single-modality approaches. This suggests that the combined use of anatomical-metabolic, inflammatory, and morphometric features captures complementary pathophysiological information, enabling more robust discrimination between non-cancerous, benign, and malignant lesions. The high accuracy achieved, particularly with the RF model, indicates that conventional ML techniques—when optimized and applied to well-curated multimodal datasets—can rival or even surpass performance levels reported in recent deep learning studies on similar diagnostic tasks. These findings are consistent with prior evidence that multimodal integration enhances predictive power in oncologic imaging, while also extending the literature by demonstrating its applicability to a three-class classification problem. From a clinical perspective, this approach has the potential to reduce unnecessary biopsies, accelerate decision-making, and support personalized diagnostic strategies in lung cancer management.

Discussion

This study demonstrated the effectiveness of a multimodal ML approach in classifying lung cancer into non-cancerous, benign, and malignant categories. By integrating sarcopenia indicators, inflammatory biomarkers, and PET/CT derived parameters, our findings provide evidence that the use of heterogeneous clinical and imaging features can substantially improve diagnostic accuracy in lung cancer assessment.

Novelty of this study lies in the integration of three distinct domains—PET/CT derived parameters, systemic inflammatory biomarkers, and sarcopenia-related functional and morphometric measures—into a single ML–based classification framework for lung cancer diagnosis. While each of these modalities has been independently investigated for their diagnostic or prognostic relevance, to the best of our knowledge, no prior work has combined all three into a multimodal model capable of simultaneously classifying non-cancerous, benign, and malignant lesions. Furthermore, by applying a feature importance–driven selection process, we demonstrate that a reduced subset of only 15 features can preserve near-optimal predictive performance, improving interpretability and clinical applicability compared to more complex, single-modality deep learning approaches reported in the literature.

Among the five tested ML algorithms, RF achieved the highest overall performance, yielding a test accuracy

Table 6 The comparison of recent literature

Study / Model	Data type / Features used	Method	Accuracy	AUC
Lv et al. [20]	CT imaging only (multi-scale fusion, deep learning)	CNN-based deep learning	94.82%	–
Abdullah et al. [21]	UCI-sourced lung cancer dataset	SVM	95.56%	–
Delzell et al. [22]	Radiomic features from CT	ML models (e.g., LR, SVM)	–	0.72
Kotoulas et al. [23]	Routine blood indices / Imaging	AI models (unspecified)	95.7 / 82.0%	–
Our Study (2025)	Multimodal: PET/CT (SUVmax, MTV), sarcopenia (SMI, gait speed), inflammatory biomarkers (CRP, ALB/BIL)	RF (with feature selection)	97%	0.99

of 96% and an average AUC of 0.99. This finding aligns with previous literature showing the robustness of ensemble learning methods in medical classification tasks involving high-dimensional, heterogeneous data [13, 14]. Furthermore, when the model was retrained using only the top 15 most informative features—dominated by PET/CT parameters (e.g., SUVmax, TLG, MTV), along with sarcopenia and inflammation-related indices (e.g., SMI, ALB/BIL)—the test accuracy improved to 97%. This result confirms that a reduced, well-curated feature set can preserve predictive power while enhancing model interpretability and efficiency. It is important to note that reported performance metrics consistently refer to the independent test set, with cross-validation used solely during the training phase.

When compared to recent artificial intelligence–based studies on lung cancer classification, our model demonstrated competitive or superior performance. For instance, Lv et al. [20] developed MIFNet, a deep learning model with multi-scale fusion for lung nodule classification, achieving 94.82% accuracy and an F1-score of 97.34% using CT imaging alone [20]. Abdullah et al. [21], using SVM on UCI-sourced lung cancer data, achieved 95.56% accuracy [21], while Delzell et al. [22] applied radiomic features from CT to traditional ML models and reached an average AUC of 0.72, which is notably lower than the 0.99 AUC obtained in our study [22]. Furthermore, [23] summarized that AI models achieved accuracies of 95.7% using routine blood indices and 82.0% with imaging to differentiate malignant from benign cases [23]. The comparison of recent literature is summarized in Table 6.

In contrast, our RF classifier—especially when restricted to the top 15 multimodal features—achieved a test accuracy of 97% and an AUC of 0.99, surpassing most previously reported values. This suggests that the combination of imaging, sarcopenia assessment, and inflammatory biomarkers within a carefully engineered ML pipeline can outperform many deep learning models relying on a single modality. The inclusion of physical function and biochemical markers likely contributed additional discriminatory power not captured by imaging alone, offering a holistic perspective on tumor biology and host status.

The importance of PET/CT derived parameters such as SUVmax, SUVort, TLG, and MTV in the classification process is consistent with prior studies that have reported their prognostic and diagnostic relevance in lung cancer staging and treatment response [7, 8]. The dominance of these imaging markers—together accounting for nearly 70% of the model's classification power—highlights the critical role of metabolic tumor burden in differentiating malignancies.

Notably, the inclusion of sarcopenia parameters such as SMI, gait speed, and albumin-based ratios contributed meaningfully to the model's decision-making. This observation supports earlier evidence that sarcopenia is not only a prognostic factor but also a potential diagnostic marker in lung cancer [10, 24]. Particularly, the inclusion of functional and M-sarcopenia assessments adds value to imaging-based models by incorporating indicators of systemic health and physical resilience, which may correlate with cancer biology.

Inflammatory biomarkers also played a non-negligible role in classification. Variables such as CRP, ALT, and composite ratios like ALB/BIL and NEU/WBC were among the top contributors, echoing findings from Guo et al. [3] and Su et al. [5], who reported the predictive utility of systemic inflammation markers in lung cancer progression and treatment response.

This study contributes to a growing body of evidence advocating for multimodal ML-based classification strategies in oncology. By integrating quantitative imaging, laboratory values, and physical performance metrics, clinicians can leverage ML models not only for diagnosis but potentially for risk stratification, follow-up planning, and treatment personalization.

However, several limitations should be acknowledged. The dataset, while diverse, was sourced from a single institution, and external validation using independent cohorts is needed to confirm generalizability. Additionally, although RF provided high accuracy and feature interpretability, future work could explore ensemble methods combining model types or incorporate explainability frameworks (e.g., SHAP, LIME) for more transparent decision support.

Future research could extend our approach by incorporating deep learning techniques. Convolutional neural networks may allow for more nuanced extraction of imaging

features, while Transformer-based architectures could integrate multimodal data streams (e.g., imaging and clinical records). However, these approaches typically require larger datasets than the one used in this study, suggesting that multi-institutional collaborations will be essential for future applications.

Conclusion

This study presents a novel multimodal ML framework that integrates PET/CT derived anatomical-metabolic parameters, inflammatory biomarkers, and sarcopenia-related variables for the classification of lung cancer into non-cancerous, benign, and malignant categories. The RF model demonstrated the highest diagnostic performance, maintaining near-perfect accuracy and AUC, particularly when applied with a reduced set of top features. The performance of this optimized model surpassed most previously published AI-based approaches that rely on single-modality data such as imaging or laboratory values alone.

These findings underscore the value of combining anatomical, functional, and biochemical markers in enhancing classification performance and clinical interpretability. In particular, our analysis of RF feature importance revealed that PET/CT derived variables contributed ~ 69% of predictive power, compared to 18% for inflammatory biomarkers and 13% for sarcopenia measures. This result represents a central contribution of our study, demonstrating both the dominant role of imaging and the complementary value of systemic and functional health indicators. Importantly, it is also shown that high accuracy can be maintained even with a reduced number of features, improving the model's efficiency and potential applicability in clinical practice.

Future studies should focus on external validation with larger, multi-institutional datasets and investigate the integration of explainable AI techniques to enhance transparency and clinician trust. Additionally, prospective applications of this model in real-time diagnostic workflows may further reveal its utility in improving patient outcomes in oncology settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13246-025-01650-x.

Author contributions Author 1 designed the study, performed machine learning analyses, and drafted the manuscript. Author 2 and 4 contributed to PET/CT data interpretation and imaging-based variable selection. Author 3 assisted with statistical validation and interpretation of results. Author 5 supported the clinical evaluation of sarcopenia and inflammatory markers. Author 6 and 7 contributed to the literature review and manuscript organization, and reviewed the final version of the manuscript and provided general feedback.

Funding This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Grant Number 123S718. The authors thank TUBITAK for their support.

Data availability The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no conflicts of interest related to this work.

Ethical approval This study was conducted in accordance with the ethical standards of Altinbas University Ethics Committee and was approved under protocol number 2022/157. Informed consent was obtained from all participants included in the study.

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A (2023) Cancer statistics, 2023. Cancer J Clin 73:17–48. https://doi.org/10.3322/caac.21763
2. Wahidi MM, Govert JA, Goudar RK, Gould MK, McCrory DC, (2007) Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer? ACCP evidence-based clinical practice guidelines (2nd edition). Chest,132:94S–107S. https://doi.org/10.1378/chest.07-1352
3. Guo D, Zhang J, Jing W, Liu J, Zhu H, Fu L et al (2018) Prognostic value of systemic immune-inflammation index in patients with advanced non-small-cell lung cancer. Future Oncol 14:2643–2650. https://doi.org/10.2217/fon-2018-0285
4. Stares M, Brown LR, Abhi D, Phillips I (2024) Prognostic biomarkers of systemic inflammation in non-small cell lung cancer: a narrative review of challenges and opportunities. Cancers (Basel) 16:1508. https://doi.org/10.3390/cancers16081508
5. Su J, Li Y, Tan S, Cheng T, Luo Y, Zhang L (2025) Pretreatment neutrophil-to-lymphocyte ratio is associated with immunotherapy efficacy in patients with advanced cancer: a systematic review and meta-analysis. Sci Rep 15:446. https://doi.org/10.1038/s41598-024-84890-3
6. AL-Jahdali H, Khan AN, Loutfi S, Al-Harbi AS (2012) Guidelines for the role of FDG-PET/CT in lung cancer management. J Infect Public Health 5:S35–40. https://doi.org/10.1016/j.jiph.2012.09.003
7. Brose A, Miederer I, König J, Gkika E, Sahlmann J, Schimek-Jasch T et al (2024) Prognostic value of metabolic tumor volume on [18F] FDG PET/CT in addition to the TNM classification system of locally advanced non-small cell lung cancer. Cancer Imaging 24:171. https://doi.org/10.1186/s40644-024-00811-7
8. Park SY, Cho A, Yu WS, Lee CY, Lee JG, Kim DJ et al (2015) Prognostic value of total lesion glycolysis by 18F-FDG PET/CT in surgically resected stage IA non-small cell lung cancer. J Nucl Med 56:45–49. https://doi.org/10.2967/jnumed.114.147561
9. Baracos VE, Martin L, Korc M, Guttridge DC, Fearon KCH (2018) Cancer-associated cachexia. Nat Rev Dis Primers 4:17105. https://doi.org/10.1038/nrdp.2017.105
10. Yang M, Shen Y, Tan L, Li W (2019) Prognostic value of sarcopenia in lung cancer: a systematic review and meta-analysis. Chest 156:101–111. https://doi.org/10.1016/j.chest.2019.04.115
11. Chianca V, Albano D, Messina C, Gitto S, Ruffo G, Guarino S et al (2022) Sarcopenia: imaging assessment and clinical application. Abdom Radiol 47:3205–3216. https://doi.org/10.1007/s00261-021-03294-3

12. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L et al (2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 25:954–961. https://doi.org/10.1038/s41591-019-0447-x
13. Gandhi Z, Gurram P, Amgai B, Lekkala SP, Lokhandwala A, Manne S et al (2023) Artificial intelligence and lung cancer: impact on improving patient outcomes. Cancers (Basel) 15:5236. https://doi.org/10.3390/cancers15215236
14. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J (2023) From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell 186:1772–1791. https://doi.org/10.1016/j.cell.2023.01.035
15. Yuan L, An L, Zhu Y, Duan C, Kong W, Jiang P et al (2024) Machine learning in diagnosis and prognosis of lung cancer by PET-CT. Cancer Manag Res 16:361–375. https://doi.org/10.2147/CMAR.S451871
16. Shah U, Biswas MR, Alzubaidi MS, Ali H, Alam T, Househ M et al (2022) Recent developments in artificial intelligence-based techniques for prostate cancer detection: a scoping review. Stud Health Technol Inf 289:268–271. https://doi.org/10.3233/SHTI210911
17. Alataş E (2025) Tanyıldızı Kökkülünk,Handan, Tanyıldızı, Hilal, and Alcın G. Treatment prediction with machine learning in prostate cancer patients. Comput Methods Biomech BioMed Eng 28:572–580. https://doi.org/10.1080/10255842.2023.2298364
18. Ahluwalia VS, Parikh RB (2025) Explainable machine learning to predict treatment response in advanced non-small cell lung cancer. JCO Clin Cancer Inf 9:e2400157. https://doi.org/10.1200/CCI-24-00157
19. Hendrix W, Hendrix N, Scholten ET, Mourits M, Trap-de Jong J, Schalekamp S et al (2023) Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Commun Med 3:1–12. https://doi.org/10.1038/s43856-023-00388-5
20. Lv E, Kang X, Wen P, Tian J, Zhang M (2024) A novel benign and malignant classification model for lung nodules based on multi-scale interleaved fusion integrated network. Sci Rep 14:27506. https://doi.org/10.1038/s41598-024-79058-y
21. Abdullah D, Mohsin Abdulazeez A, Sallow A (2021) Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Acad J 1:141–149. https://doi.org/10.48161/qaj.v1n2a58
22. Delzell DAP, Magnuson S, Peter T, Smith M, Smith BJ Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. Front Oncol 2019; 9:1393.
23. Kotoulas S-C, Spyratos D, Porpodis K, Domvri K, Boutou A, Kaimakamis E et al (2025) A thorough review of the clinical applications of artificial intelligence in lung cancer. Cancers 17:882. https://doi.org/10.3390/cancers17050882
24. Dong X, Dan X, Yawen A, Haibo X, Huan L, Mengqi T et al (2020) Identifying sarcopenia in advanced non-small cell lung cancer patients using skeletal muscle CT radiomics and machine learning. Thorac Cancer 11:2650–2659. https://doi.org/10.1111/1759-7714.13598
25. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11):e0224365. https://doi.org/10.1371/journal.pone.0224365PMID: 31697686; PMCID: PMC6837442
26. Deo RC (2015) Machine learning in medicine. Circulation 132(20):1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593PMID: 26572668; PMCID: PMC5831252
27. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression (3rd ed.). Wiley. New Jersey.
28. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
29. Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI, Iftekharuddin KM (2019) Feature-Guided deep radiomics for glioblastoma patient survival prediction. Front Neurosci 13:966. https://doi.org/10.3389/fnins.2019.00966PMID: 31619949; PMCID: PMC6763591
30. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
31. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
32. Belgiu M, Drăguţ L.Random forest in remote sensing: a review of applications and future directions,ISPRS. J Photogramm Remote Sens, (2016); 114:24–31, ISSN 0924–2716. https://doi.org/10.1016/j.isprsjprs.2016.01.011
33. Breiman L, Random Forests (2001) Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Authors and Affiliations

Handan Tanyildizi-Kokkulunk[1] · Goksel Alcin[2] · Iffet Cavdar[3] · Resit Akyel[4] · Safak Yigit[5] · Tuba Ciftci-Kusbeci[6] · Gonul Caliskan[7]

✉ Handan Tanyildizi-Kokkulunk
handan.kokkulunk@altinbas.edu.tr

Goksel Alcin
gokselalcin@hotmail.com.tr

Iffet Cavdar
icavdar@istanbul.edu.tr

Resit Akyel
akyelresit@hotmail.com

Safak Yigit
safak.yigit@galata.edu.tr

Tuba Ciftci-Kusbeci
ciftcituba@gmail.com

Gonul Caliskan
gonul.caliskan@bilgiedu.net

[1] Radiotherapy Program, Vocational School of Health Sciences, Altınbaş University, Kartaltepe Dist., No.11, Bakirkoy, 34147 Istanbul, Turkey

[2] Department of Nuclear Medicine, Istanbul Training and Research Hospital, Cerrahpasa, Org. Abdurrahman Nafiz Gurman Cd. No:24, Fatih, 34098 Istanbul, Turkey

[3] Department of Nuclear Physics, Faculty of Science, Istanbul University, Vezneciler, Fatih, 34134 Istanbul, Turkey

[4] Department of Nuclear Medicine, Yedikule Chest Diseases Hospital, Kazlıcesme, Zeytinburnu, 34020 Istanbul, Turkey

[5] Physiotherapy Program, Vocational School, Istanbul Galata University, Evliya Celebi Dist., Mesrutiyet St., No:62, Beyoglu, 34430 Istanbul, Turkey

[6] Department of Chest Diseases, Faculty of Medicine, Altınbaş University, Bahcelievler MedicalPark Hospital, E-5 Highway, Kultur St., No:1, Bahçelievler, 34147 Istanbul, Turkey

[7] Physiotherapy and Rehabilitation Master Program, Institute of Graduate Programs, Istanbul Bilgi University, Eyüpsultan, 34060 Istanbul, Turkey