CMJ.2025;47(2): 35-44
DOİ: 10.7197/cmj.1713462

Cumhuriyet Medical Journal

| cmj.cumhuriyet.edu.tr | Founded: 2004 Available online, ISSN:1305-0028 Publisher: Sivas Cumhuriyet Üniversitesi

Development Of An Artificial İntelligence-Based Precision Medicine Decision Support System For Radiogenomics Data Sets

Abdulvahap Pınar[1,a,*], Ahmet Kadir Arslan[1,b], Emek Güldoğan[1,c]

[1] *Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya 44280, Turkey*

**Corresponding author*

Research Article

History

Received: *03/06/2025*
Accepted: *11/06/2025*

ABSTRACT

Aim: This study aims to apply deep learning algorithms for superpixel segmentation, herbaceous thresholding, and disease reference position estimation from DICOM images and clinical data of Non-Small Cell Lung Cancer (NSCLC) patients. Quantitative imaging data was integrated with clinical information. Various machine learning algorithms were employed to identify biomarkers and evaluate classification performance based on clinical data, imaging data, and their combination, assessing the model improvement rates.

Materials and Methods: The clinical dataset included 43 patients with and 168 without an Epidermal Growth Factor Receptor (EGFR) mutation, and 38 with and 173 without a Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS) mutation, totaling 211 NSCLC cases. A total of 2,231 images were analyzed. Using the VGG16 deep learning model, 25,088 features were extracted from each image. XGBoost, CatBoost, Random Forest, and Support Vector Machine (SVM) classification algorithms were used to predict mutation status.

Results: Clinical data revealed significant differences in mutation status among NSCLC patients. The Random Forest algorithm was employed for feature selection, identifying the 50 most important variables for model training. XGBoost and CatBoost achieved the highest classification performance, with results for accuracy, balanced accuracy, precision, sensitivity, F1-score, and ROC-AUC as follows: 0.965 ± 0.015, 0.954 ± 0.021, 0.953 ± 0.024, 0.994 ± 0.007, 0.973 ± 0.011, and 0.990 ± 0.005, respectively.

Conclusion: The study's findings demonstrate that XGBoost and CatBoost models were highly effective in predicting KRAS mutation status from imaging data. CatBoost also performed best in determining EGFR mutation status, outperforming other machine learning methods.

Keywords: Lung cancer, Biomedical, Digital imaging, Machine learning

Radyogenomik Veri Setleri İçin Yapay Zeka Tabanlı Bir Hassas Tıp Karar Destek Sisteminin Geliştirilmesi

Araştırma Makalesi

Süreç

Geliş: *03/06/2025*
Kabul: *11/06/2025*

ÖZ

Amaç: Bu çalışma, küçük hücreli dışı akciğer kanseri (KHDAK) hastalarına ait DICOM görüntüleri ve klinik verilerden süperpiksel segmentasyonu, otsu eşikleme ve hastalık referans pozisyonu tahmini için derin öğrenme algoritmalarını uygulamayı amaçlamaktadır. Nicel görüntüleme verileri, klinik bilgilerle entegre edilmiştir. Klinik veriler, görüntüleme verileri ve bunların kombinasyonuna dayalı olarak biyobelirteçleri tanımlamak ve sınıflandırma performansını değerlendirmek için çeşitli makine öğrenmesi algoritmaları kullanılmış; model iyileşme oranları değerlendirilmiştir.

Gereç ve Yöntem: Klinik veri seti, Epidermal Büyüme Faktörü Reseptör (EGFR) mutasyonu olan 43 ve olmayan 168, Kirsten Rat Sarkom Viral Onkogen Homoloğu (KRAS) mutasyonu olan 38 ve olmayan 173 hasta olmak üzere toplam 211 KHDAK vakasını içermektedir. Toplam 2.231 görüntü analiz edilmiştir. VGG16 derin öğrenme modeli kullanılarak her bir görüntüden 25.088 özellik çıkarılmıştır. Mutasyon durumunu tahmin etmek için XGBoost, CatBoost, Random Forest ve Destek Vektör Makineleri (SVM) sınıflandırma algoritmaları kullanılmıştır.

Bulgular: Klinik veriler, KHDAK hastaları arasında mutasyon durumlarına göre anlamlı farklılıklar olduğunu ortaya koymuştur. Model eğitimi için en önemli 50 değişkeni belirlemek amacıyla Random Forest algoritması ile özellik seçimi yapılmıştır. XGBoost ve CatBoost, en yüksek sınıflandırma performansını elde etmiştir. Elde edilen doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1 skoru ve ROC-AUC değerleri sırasıyla şu şekildedir: 0.965 ± 0.015, 0.954 ± 0.021, 0.953 ± 0.024, 0.994 ± 0.007, 0.973 ± 0.011 ve 0.990 ± 0.005.

Sonuç: Çalışmanın bulguları, XGBoost ve CatBoost modellerinin görüntüleme verilerinden KRAS mutasyon durumunu tahmin etmede son derece etkili olduğunu göstermektedir. Ayrıca CatBoost, EGFR mutasyon durumunun belirlenmesinde de diğer makine öğrenmesi yöntemlerinden daha iyi performans göstermiştir.

Anahtar Kelimeler: Akciğer kanseri, Biyomedikal, Dijital görüntüleme, Makine öğrenimi

[a] *apinar@adiyaman.edu.tr* *0000-0002-3662-2579*
[c] *emek.guldogan@inonu.edu.tr* *0000-0002-5436-8164*
[b] *arslan.ahmet@inonu.edu.tr* *0000-0001-8626-9542*

How to Cite: Pınar A, Arslan AK, Guldogan E. Development of an Artificial İntelligence-Based Precision Medicine Decision Support System for Radiogenomics Data Sets, Cumhuriyet Medical Journal. 2025;47(2): 35-44

Introduction

Since the genomics revolution in the early 1990s, cancer research has concentrated on determining the genetic origins of illnesses in order to allow precision medicine therapies. Following the completion of the human genome project, the design of genes and proteins has progressed to functional levels of gene science and gene functions. Many cancer studies, such as the Cancer Genome Atlas (TCGA), have tried to acquire transcriptome, epigenomic, and proteomic data on genome type. [1,2] Consequently, protocols addressing the primary hazards and potential aggravating symptoms associated with genetic analysis technologies have been established.

Genetic analysis may fail to correctly represent genetic variations in tissue biopsy samples due to intra- and inter-group variability of tumor variables.[3] As precision medicine and big data research advance, several professionals in the area underscore the significance of "Radiogenomics". Radiogenomics facilitates the establishment of multi-scale correlations between medical imaging and genetic data, enhancing analogous linkages and integrating the overlooked elements of radiomics and genomics. [4] Radiogenomics is a commonly utilized approach for predicting gamete mutations known as phenotype or genotype. To apply the classification model, the radiogenomic structure is divided into two or more groups, and the classification models are used to generate the appropriate prediction model.

Model prediction for radiogenomics categorization is typically real-valued. The anticipated performance values are determined by considering true positive, true negative, false positive, and false negative values. [5-6] Lung cancer-related deaths are highly widespread over the World.[7] Medical imaging radiography or computed tomography (CT) is used to diagnose lung cancer, and the results are often the existence of a lesion in the lung and the interaction between this tumor and the surrounding tissues. The discovered lesions are often biopsied to determine the cancer diagnosis and histological symptoms of the tumor, such as small cell lung cancer (SCLC), non-small cell lung cancer (NSCLC), and so on. [7]

Deep Learning (DL) models are a valuable tool in medical image analysis. Models derived from deep learning network architectures are extremely useful in many fields of health sciences, particularly medical image identification and segmentation. It is commonly used in areas such as early diagnosis and therapy. [8] DL challenges advance at a slower rate than other real-world problems in medical imaging and medical services. By evaluating the variables influencing the rise in the development of DL architectures and in accordance with medical imaging investigations, the application fields of Computed Tomography (CT) technology were explored and the associated topics were highlighted. [9]

The fundamental method for detecting the radiogenomics of lung cancer using DL techniques involves assessing the alterations in genomic biomarkers obtained from CT scans. A convolutional neural network (CNN) model was developed to analyze the epidermal growth factor (cell division, cell growth, cell survival, etc.) for assessing the mutation status using CT devices and identifying the most effective tree-based technique for the ideal procedure.[10] Image analysis has demonstrated remarkable success employing CNNs in the medical industry for learning feature detection in lung nodes and mass segmentation applications. [11]

Although databases designed to compare data on publicly available biomedical images help to develop image analysis algorithms by providing resources for users to evaluate and compare previous models as well as generate new models, the fact that some datasets are distributed in multiple locations or indexed using different terminologies makes reliable model comparison and reproducibility difficult. The Lung Image Database Consortium (LIDC) image collection allows for a comparison of the fundamental accuracy of biological datasets to models.[12-13]

This study aims to develop a medical decision support system that preprocesses clinical and radiogenomic datasets for NSCLC and then builds Machine Learning (ML) based prediction models for lung cancer.

Materials and Methods

Dataset

This dataset comprises detailed medical records and imaging data from 211 NSCLC patients. The collection includes both well-reviewed descriptions of malignant tumors apparent on medical imaging and quantitative data on the related CT scan images. Tumor segmentation pictures from PET/CT scans are also presented.[14] The dataset included 211 samples, with 135 (64%) men and 76 (36%) females. There were 19 (14.10%) male patients having an Epidermal Growth Factor Receptor (EGFR) mutation. Male patients without EGFR mutation were 116 (85.90%), while male patients with Kirsten Rat Sarcoma viral oncogenic homolog (KRAS) mutation were 27 (20%) and male patients without KRAS mutation were 108 (80%). There were 24 (31.60%) female patients having an EGFR mutation. There were 52 female patients (68.40%) who did not have EGFR mutation. Female patients with KRAS mutation were 11 (14.50%), whereas female patients without KRAS mutation were 65 (85.50%).

Radiogenomics

Radiogenomics, which is developed by combining "Radiomics" and "Genomics" in the field of artificial intelligence(AI) health, has gained its position in the literature as the most recent technology science employed in the disciplines of precision medicine and cancer, as well as in other departments of science. Radiogenomics is the categorization of risk that combines precision medicine genetic data with large volumes of radiographic imaging data. Many AI studies involving patients in clinical settings have been used to develop models. In oncology research, life analytic forecasts and studies have demonstrated significant accomplishments using radiogenomic outcomes facilitated by AI.[15]

Substantial advancements have been achieved in lung cancer treatment, encompassing sophisticated screening techniques employed by specialists in conjunction with artificial intelligence, the implementation of minimally invasive diagnostic and therapeutic procedures, radiation modalities such as stereotactic ablative therapy, and the development of novel targeted therapies and immunotherapeutics.[16] The introduction of these novel therapy modalities has been linked to enhanced survival rates, especially in patients with non-small cell lung cancer. The two-year relative survival rate for NSCLC rose from 34% in 2009-2010 to 42% in 2015-2016. [17]

Image Processing

Image processing techniques are employed to segment lung tissue and lesions. This procedure seeks to retrieve useful information by detecting and assessing the borders of lesions. Researchers underline the efficiency of this strategy by investigating the performance of automated lesion detection. [18] Lesions are classified using ML algorithms based on their shape and density. This approach is proposed to distinguish between malignant and precancerous lesions.[19] The outcomes of automated lesion recognition in CT scans were evaluated using sensitivity, specificity, and other metrics from image processing performance measurements. [20]

Artificial Intelligence (AI)

Machines can use AI algorithms to compare data, arrange patient follow-up in the health sector, learn data, communicate, perceive, interpret images, and move and displace items. It is also a science that aids physicians in their decision-making processes while diagnosing and treating patients, as well as diagnosing diseases. AI may be used to create systems that imitate certain human actions (such as picking up things and depositing them in specified areas) as well as human thought processes (such as data computation and medical diagnosis).

Although substantial improvements have been achieved, there is currently very little study in the subject of AI, and AI researchers are continually developing new technologies. The basic methodologies of AI include fuzzy logic, artificial neural networks, genetic algorithms, and expert systems. Computer systems can create, plan, diagnose, interpret, summarise, generalize, control, and provide suggestions [21].

DL techniques suggest that some illnesses can be identified using radiological data. Chest radiography, for example, is the most prevalent form of radiological examination in the world, with a vast dataset. DL models appear to be able to identify clinically significant anomalies in chest radiography such as pneumonia and pneumothorax [22].

Machine Learning (ML)

A machine learning (ML) algorithm is a subfield of artificial intelligence that investigates strategies for improving data set performance by developing skill in processing large amounts of data. ML explains the data set using Supervised, Unsupervised, Semi-supervised, and Reinforcement learning approaches based on the data set's output variables. It gives the researcher a broad variety of information regarding the correlations between input and output data. ML algorithms are capable of solving a variety of perceptual problems.

The basic objective of ML is to create predictive models capable of making data-driven assessments and choices while producing accurate and consistent forecasts. These models may be used to a broad range of tasks, including image recognition, model accuracy, natural language processing, and fraud detection. Its rising prominence may be attributed to the availability of vast volumes of data and increases in computer processing capacity [23,24].

Approaches to Preprocessing Clinical Data

Data mining and modeling is the process of preparing data and increasing data quality in order to make the data processing more efficient and accurate. This procedure often includes data cleaning, data transformation, data standardization, data reduction, RF and variable selection, and other activities necessary to prepare the data for future analysis. These stages fill in data gaps, fix discrepancies, and prepare the data for analysis [25]. This study employed “scikit-learn” for categorical data and utilized “TensorFlow” and “keras” libraries for picture normalization, since the “VGG16” model offers superior performance in data transformation.

Preprocessing Approaches for Image Data

Image preprocessing refers to a collection of tools and approaches used to analyze and meaningfully transform the raw data of digital photographs. This technique consists of multiple processes to increase picture quality, remove noise, and highlight certain characteristics.Image processing steps that make a picture suitable for analysis include adjusting the brightness, contrast, and color balance; cropping out unwanted areas; resizing the picture to make it a different size; applying filters to make the picture less noisy; identifying objects in the picture using edge detection methods; and finally, applying histogram equalization to make the picture more contrasty [26]. In this work, "OpenCV" was used for image normalization, "SciPy" from the python library for imge segmentation, and "OpenCV" was utilized for contour detection. Additionally, Image Thresholding Methods were employed.Additionally, Image Thresholding Methods were used.

Random Forest (RF)

Random Forest(RF) is one of the disciplines of research where it has been widely applied to image categorization. RF is recognized for its efficacy on datasets with numerous characteristics, has robust noise resilience, and attains elevated classification accuracy. The final classification outcome is determined by the majority vote of the decision trees inside the Random Forest model, which synthesizes the results from several trees trained on the data [27]. It is trained less frequently due to its enhancement of the RF algorithm's dependability and stability. The technique is highly resilient regarding generalization and model correctness, therefore offering a dependable solution for prediction and classification challenges [28]. The decision tree framework for the RF method, which generates a decision tree for each segment after partitioning the dataset into random subsets, is illustrated [29].

Extreme Gradient Boosting (XGBoost)

Chen and Guestrin created this effective ML algorithm for regression and classification procedures. It performs quite well, particularly on structured data. XGBoost is an improved version of the gradient boosting technique. This approach works by successively merging weak learners to fix model flaws, increasing the efficiency of these operations. XGBoost is widely employed across a variety of industries, including healthcare, finance, and clinical imaging. It works well with huge data sets and data with missing cells [30]. The "XGBClassifier" package for the XGBoost method was used in this work, and classification performance for models with 200 iterations was achieved.

Categorical Boosting Algorithm (CatBoost)

The CatBoost algorithm is a machine learning system capable of processing both numerical and categorical input. CatBoost mitigates overfitting, a characteristic aspect of this technique, by including the algorithm's prior values with low-frequency characteristics and regions of high density, so effectively handling points in noise. Gradient-assisted decision trees underpin the creation of the CatBoost methodology. CatBoost mitigates the bias in predicted values generated by the gradient descent method, hence enhancing data comprehension and outcome evaluation [31,32]. This study employed the "CatBoostClassifier" library for the CatBoost method, utilizing 200 iterations to get metrics on classification performance.

Support Vector Machines

The Support Vector Machines (SVM) method seeks to reduce the empirical error probability of traditional pattern recognition approaches by improving performance on the training dataset. SVM, on the other hand, is concerned with minimizing structural risk, which refers to the danger of inaccurately categorizing unseen patterns based on the data's fixed but unknown probability distribution. The concept of uniform convergence in probability led to the development of a new principle of induction, which aims to minimize an upper constraint on the generalization error [33].

Biostatistical Analyses

This study presented quantitative data using mean ± standard deviation and qualitative data using number (%). Prior to deriving conclusions from the data analysis, examinations were performed to detect absent values or severe outliers within the dataset. The data set was amended using suitable procedures, if required. The Shapiro-Wilk test was employed to assess conformity with the normal distribution assumption, hence informing the selection of hypothesis tests for data analysis. The statistical significance threshold was established at p<0.05. IBM SPSS Statistics for Windows Version 27.0 package program was used for statistical analysis [34].

Python programming language and "tensorflow, keras, preprocessing.image", "seaborn", "pandas", "OpenCv", "traceback", "os", "pydicom", "sklearn" for ML methods used in classification of biomedical data and image analysis.ensemble", "StratifiedKFold", "tqdm", "sklearn.metrics", "tensorflow, keras", "torch", "skimage.segmentation", "glob" libraries were used.

Result

The dataset for this study includes information on the presence and absence of EGFR and KRAS mutations. With 211 samples total, there are 135 males and 76 females, or 64% and 36% of the total, respectively. There were 19 male patients (14.10% of the total) who showed evidence of EGFR mutations. A total of 108 male patients (80%) lacked the KRAS mutation, 27 male patients (20%) possessed the EGFR mutation, and 116 male patients (85.90%) lacked the KRAS mutation. Among the patients, 24 had EGFR mutations, accounting for 31.60% of the total. Of the individuals analyzed, 52 were female and accounted for 68.40% without an EGFR mutation. We found 65 female patients (85.50%) without KRAS mutation and 11 female patients (14.50%) with KRAS mutation. [35]

The demographic characteristics of the categorical variables are presented in Table 1. When the descriptive statistics table of categorical variables related to EGFR mutation status was analyzed, According to Table 1, the variables of gender p=0.002, smoking status p<0.001, histology p=0.002, histopathologic grade p=0.014, and patient survival status p=0.046 (p<0.05) show a statistically significant difference between the groups in terms of EGFR mutation. In patients with EGFR mutation, 81.4% were connected with Stanford Health System/Hospital, and the proportion of female patients was 55.8%, which was greater than the proportion of male patients (44.20%).

The percentage of patients who had never smoked was 51.2%. Furthermore, all patients with the EGFR mutation exhibited adenocarcinoma histology, and their survival percentage was assessed to be 83.7%. However, there was no statistically significant difference in pathologic staging (T, N, M), lymphovascular invasion, pleural invasion, or adjuvant therapy characteristics (p>0.05). When analyzing the table based on KRAS mutation status, only the histology parameter showed a significant difference (p<0.05; p=0.010). Although 97.4% of patients with the KRAS mutation had adenocarcinoma histology, there was no statistically significant difference in all other demographic and clinicopathologic parameters (gender, ethnicity, smoking status, pathologic staging, histopathologic grade, lymphovascular invasion, pleural invasion, adjuvant treatment, and survival) (p > 0.05).

When Table 2 is analyzed according to EGFR and KRAS mutation status, there is no statistically significant difference between the groups in the variables of age at histologic diagnosis (68±10, 68±10, p=0.924) and (66±10, 68±10, p=0.165), weight (122±18, 123±22, p=0.783), The days between CT and surgery (39±27, 53±68, p=0.205) (p>0.05). The "Pack Years" variable quantifies the cumulative smoking exposure of patients, determined by multiplying the daily cigarette pack consumption by the total number of years smoked.

Table 3 provides a comparative overview of the classification performance across different models for detecting EGFR and KRAS mutations. Among these, CatBoost emerged as the top-performing algorithm in both categories. Specifically, it achieved an accuracy of 96.7%, sensitivity of 99.1%, F1-score of 97.9%, and ROC-AUC of 98.9% for EGFR, while delivering similarly high metrics for KRAS with 96.5% accuracy and 99.4% sensitivity. Although XGBoost followed closely behind, its sensitivity and F1-score remained marginally lower. Random Forest and SVM, on the other hand, yielded comparatively suboptimal outcomes, especially in balanced accuracy and precision. These results point to CatBoost’s strong predictive capacity and its potential utility in accurately distinguishing mutation types in genomic classification tasks.

Table 1. Descriptive Statistics of Categorical Variables Related To EGFR And KRAS Mutation Status

Variables	**Categories**	**EGFR Mutation**			**KRAS Mutation**		
		Yes	**No**	**p**	**Yes**	**No**	**p-value**
		N(%)	**N(%)**	0.558**	**N(%)**	**N(%)**	
Recurrence	No	30 (69.80)	127 (75.60)		28 (73.70)	129 (74.60)	0.987*
	Yes	13 (30.20)	41 (24.40)		10 (26.30)	44 (25.40%)	
Patient affiliation	Veterans Affairs System	8 (18.60)	85 (50.60)	**<0.001***	19 (50.00)	74 (42.80%)	0.417*
	Stanford Health System/Hospital	35 (81.40)	83 (49.40)		19 (50.00)	99 (57.20)	
Gender	Male	19 (44.20)	116 (69.00)	**0.002***	27 (71.10)	108 (62.40)	0.414*
	Female	24 (55.80)	52 (31.00)		11 (28.90)	65 (37.60)	
Ethnicity	Caucasian	32 (74.40)	140 (83.30)	0.552**	33 (86.80)	139 (80.30)	0.508**
	Native Hawaiian/Pacific Islander	0 (0.00)	3 (1.80)		0 (0.00)	3 (1.70%)	
	Asia	10 (23.30)	14 (8.30)		2 (5.30)	22 (12.70)	
	Afro-Amerikan	0 (0.00)	6 (3.60)		2 (5.30)	4 (2.30)	
	Hispanik/Latino	1 (2.30)	5 (3.00)		1 (2.60)	5 (2.90)	
Smoking status	Smokers	2 (4.70)	31 (18.50)	**<0.001***	8 (21.10)	25 (14.50)	0.119**
	Previous users	19 (44.20)	111 (66.10)		26 (68.40)	104 (60.10)	
	Non-smokers	22 (51.20)	26 (15.50)		4 (10.50)	44 (25.40)	
Histology	Adenokarsinom	43 (100.00)	129 (76.80)	**0.002****	37 (97.40)	135 (78.00)	**0.010****
	Squamous cell carcinoma	0 (0.00)	35 (20.80)		0 (0.00)	35 (20.20)	
Pathological Tumor stage	T1a	2 (4.70)	38 (22.60)	0.080**	8 (21.10)	32 (18.50)	0.750**
	T1b	8 (18.60)	23 (13.70)		5 (13.20)	26 (15.00)	
	T2a	27 (62.80)	69 (41.10)		16 (42.10)	80 (46.20)	
	T2b	1 (2.30)	9 (5.40)		2 (5.30)	8 (4.60)	
	T3	3 (7.00)	18 (10.70)		6 (15.80)	15 (8.70)	
	T4	1 (2.30)	6 (3.60)		0 (0.00)	7 (4.00)	
	Tis	1 (2.30)	5 (3.00)		1 (2.60)	5 (2.90)	
Pathologic Lymph Node stage	N0	38 (88.40)	140 (83.30)	0.699**	32 (84.20)	146 (84.40)	0.969**
	N1	2 (4.70)	13 (7.70)		3 (7.90)	12 (6.90)	
	N2	3 (7.00)	15 (8.90)		3 (7.90)	15 (8.70)	
Pathologic Metastasis stage	M0	42 (97.70)	164 (97.60)	0.989**	36 (94.70)	170 (98.30)	0.221**
	M1b	1 (2.30)	4 (2.40)		2 (5.30)	3 (1.70)	
Histopathological Grades	G1	7 (16.30)	25 (14.90)	**0.014****	5 (13.20)	27 (15.60)	0.650**
	G2	29 (67.40)	96 (57.10)		21 (55.30)	104 (60.10)	
	G3	0 (0.00)	33 (19.60)		9 (23.70)	24 (13.90)	
	Other, Type I	4 (9.30)	5 (3.00)		1 (2.60)	8 (4.60)	
	Other, Type II	3 (7.00)	9 (5.40)		2 (5.30)	10 (5.80)	
Lymphovascular invasion	Yes	3 (7.00)	18 (10.70)	0.579*	4 (10.50)	17 (9.80)	0.968*
	No	40 (93.00)	150 (89.30)		34 (89.50)	156 (90.20)	
Pleural invasion (elastic, visceral or parietal)	No	34 (79.10)	135 (80.40)	0.998*	30 (78.90)	139 (80.30)	0.978*
	Yes	9 (20.90)	33 (19.60)		8 (21.10)	34 (19.70)	
Adjuvant Treatment	No	37 (86.00)	125 (74.40)	0.158*	28 (73.70)	134 (77.50)	0.774*
	Yes	6 (14.00)	43 (25.60)		10 (26.30)	39 (22.50)	
Kemoterapi	No	37 (86.00)	125 (74.40)	0.158*	28 (73.70)	134 (77.50)	0.774*
	Yes	6 (14.00)	43 (25.60)		10 (26.30)	39 (22.50)	
Radiation	No	41 (95.30)	154 (91.70)	0.535*	35 (92.10)	160 (92.50)	0.980*
	Yes	2 (4.70)	14 (8.30)		3 (7.90)	13 (7.50)	
Survival Status	Dead	7 (16.30)	56 (33.30)	**0.046***	10 (26.30)	53 (30.60)	0.741*
	Live	36 (83.70)	112 (66.70)		28 (73.70)	120 (69.40)	

GFR: Epidermal Growth Factor Receptor; KRAS: Kirsten Rat Sarcoma viral oncogene homolog; G1: Well Differentiated; G2: Moderately Differentiated; G3: Poorly Differentiated; Type I: Good to moderately differentiated; Type II: Moderately to poorly differentiated; * : Pearson Chi-square; **: Fisher's Exact Test

Table 2. Descriptive Statistics For Quantitative Data

Variables			Age at Histologic Diagnosis	Weight (lbs)	Pack Years	The days between CT and surgery
EGFR Mutation Status	Yes	Mean ± SD	68±10	122±18	32±16	39±27
	No	Mean ± SD	68±10	123±22	41±24	53±68
		p*	0.924	0.783	0.004	0.209
KRAS Mutation Status	Yes	Mean ± SD	66±10	128±26	43±28	48±39
	No	Mean ± SD	68±10	121±20	38±21	51±67
		p*	0.165	0.096	0.257	0.789

EGFR: Epidermal Growth Factor Receptor; KRAS: Kirsten Rat Sarcoma viral oncogene homolog; * : Independent Two Sample t-test; SD: Standard Deviation.

Table 3. Metrics on classification performance

Group	Model	Accuracy	Balanced Accuracy	Precision	Sensitivity	F1-Score	ROC-AUC
EGFR	CatBoost	0.967 ± 0.010 (0.966 - 0.967)	0.931 ± 0.022 (0.931 - 0.932)	0.968 ± 0.010 (0.968 - 0.969)	0.991 ± 0.003 (0.990 - 0.991)	0.979 ± 0.006 (0.979 - 0.980)	0.989 ± 0.002 (0.988 - 0.989)
	Random Forest	0.888 ± 0.017 (0.887 - 0.888)	0.786 ± 0.038 (0.785 - 0.788)	0.907 ± 0.017 (0.907 - 0.908)	0.957 ± 0.006 (0.956 - 0.957)	0.931 ± 0.010 (0.931 - 0.932)	0.933 ± 0.021 (0.932 - 0.934)
	SVM	0.730 ± 0.015 (0.730 - 0.731)	0.714 ± 0.015 (0.713 - 0.714)	0.903 ± 0.011 (0.902 - 0.903)	0.742 ± 0.028 (0.741 - 0.743)	0.814 ± 0.014 (0.813 - 0.815)	0.786 ± 0.017 (0.786 - 0.787)
	XGBoost	0.963 ± 0.010 (0.963 - 0.964)	0.931 ± 0.012 (0.930 - 0.931)	0.969 ± 0.004 (0.968 - 0.969)	0.985 ± 0.009 (0.985 - 0.986)	0.977 ± 0.006 (0.976 - 0.977)	0.985 ± 0.007 (0.984 - 0.985)
KRAS	CatBoost	0.965 ± 0.015 (0.964 - 0.966)	0.954 ± 0.021 (0.953 - 0.955)	0.953 ± 0.024 (0.952 - 0.954)	0.994 ± 0.007 (0.994 - 0.995)	0.973 ± 0.011 (0.973 - 0.974)	0.990 ± 0.005 (0.990 - 0.991)
	Random Forest	0.879 ± 0.008 (0.878 - 0.879)	0.839 ± 0.010 (0.839 - 0.840)	0.848 ± 0.012 (0.848 - 0.849)	0.986 ± 0.020 (0.985 - 0.987)	0.912 ± 0.007 (0.911 - 0.912)	0.942 ± 0.008 (0.942 - 0.942)
	SVM	0.813 ± 0.024 (0.812 - 0.814)	0.795 ± 0.018 (0.794 - 0.796)	0.846 ± 0.010 (0.845 - 0.846)	0.863 ± 0.043 (0.861 - 0.864)	0.854 ± 0.022 (0.853 - 0.855)	0.847 ± 0.013 (0.847 - 0.848)
	XGBoost	0.951 ± 0.006 (0.950 - 0.951)	0.936 ± 0.007 (0.935 - 0.936)	0.935 ± 0.009 (0.934 - 0.935)	0.991 ± 0.011 (0.991 - 0.992)	0.962 ± 0.005 (0.962 - 0.962)	0.982 ± 0.011 (0.981 - 0.982)

Figure 1 depicts the superpixel segmentation approach applied to lung slices from an NSCLC patient. By isolating the tumor from its surroundings, this technique serves as a critical preprocessing step for visual analysis and ML-based diagnostic systems.

The Otsu thresholding approach, seen in Figure 2, facilitated the delineation of lung structures from the background and effectively highlighted lung areas throughout the segmentation phase.

Figure 3 illustrates (a) Grid: a geometric network framework employed for superpixel segmentation; (b) the original grayscale DICOM image; (c) the binary segmentation mask delineating the tumor region in white; and (d) the analysis of the final reference points, with boundaries defined in red, alongside the evaluation and classification data.

Figure 4 shows that the tumor locations detected from the pictures were colored green. Tumor centers were delineated in blue, and contours were used to identify tumor borders. Contours were delineated using red lines. In the NSCLC image processing phase, (a) Grid: denotes a geometric network framework utilized for superpixel segmentation; (b) the original grayscale DICOM image; (c) the binary segmentation mask illustrating the tumor region in white; and (d) signifies the final analytical outcome, encompassing a color image alongside the evaluation and classification data.

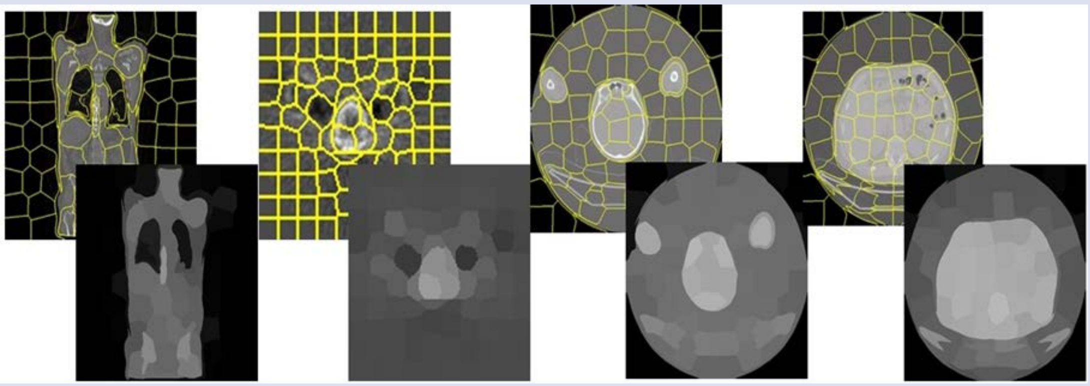

Figure 1. Superpixel Segmentation İn CT İmages And İts Results

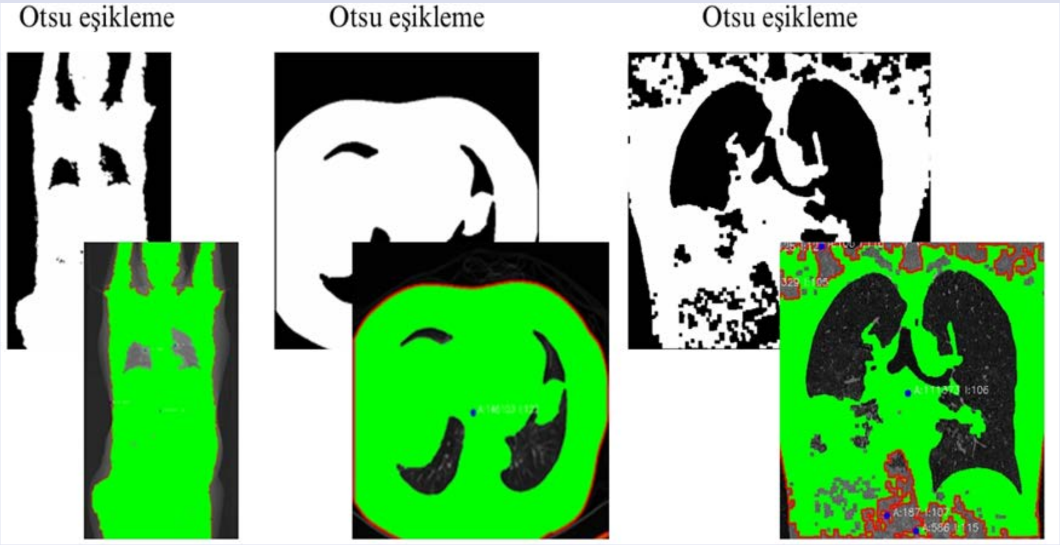

Figure 2. Masking The İmage After Herbaceous Thresholding And Determination Of Reference Points

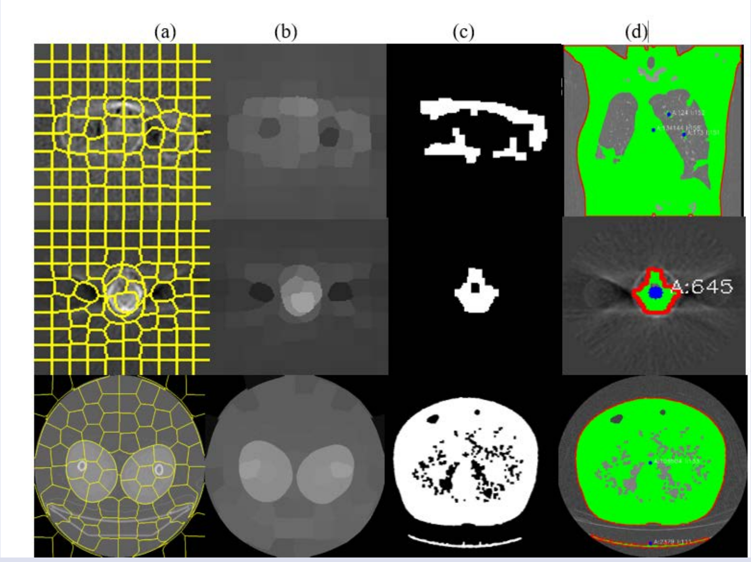

Figure 3. Segmentation, Masking And Contour Analysis Of DICOM İmages Of NSCLC Patients

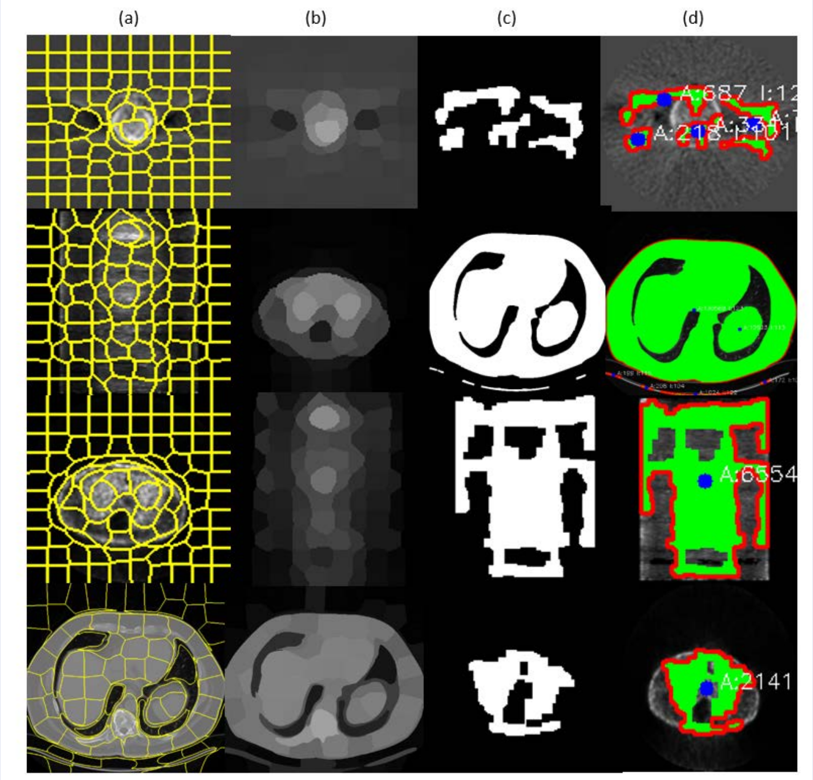

Figure 4. Segmentation, Masking And Contour Analysis İn Lung İmages.

Discussion

The structural characteristics of NSCLC have lately provided professionals with an edge in patient therapy and decision-making. Mutations in EGFR and KRAS enable healthcare providers to enhance therapeutic targets and implement tailored therapy. The invasiveness of repeated biologics and tumor heterogeneity provide significant clinical problems in genomic profiling. Radiomics utilizes several features to extract picture characteristics from high-throughput radiographic images, facilitating biomarker prediction and non-invasive estimation of lesion phenotypes. Extensive research has been conducted on the use of radiomic characteristics in lung cancer prevention, including the selection of tumor phenotypes and the prediction of biomarkers. Non-invasive imaging modalities, like ultrasound and MRI, are used to diagnose conditions such as cancer [36].

Imaging methods are used by healthcare practitioners to identify disease types and forecast illness progression, facilitating early diagnosis and treatment. Non-invasive imaging technologies (CT, MRI, PET) are used to ascertain tumor size, quantity, and density within an image. Imaging techniques not only identify the existence of pulmonary illness but also provide critical information on biological or clinical biomarkers, including prognostic indicators such as disease progression, treatment efficacy, and life expectancy [37].

This aim of the study was to assess the prevalence of EGFR and KRAS mutations in NSCLC patients in relation to clinical data and to analyze the effectiveness of different machine learning methodologies for early detection of these mutations in DICOM images. It may facilitate the implementation of customized medicine and targeted therapeutic choices. The research on The Cancer Imaging Archive (TCIA) comprises a comprehensive clinical dataset of 211 individuals with non-small cell lung cancer (NSCLC) and biomedical imaging datasets [38,39]. A study utilizing the same dataset, augmented by 161 cohort patients from a total of 211 NSCLC patients, revealed that the XGBoost model achieved EGFR and KRAS scores of 0.83 and 0.86, respectively, through 10-fold cross-validation for predicting EGFR and KRAS mutations. The AUC-ROC values for these mutations were recorded at 0.89 and 0.812, respectively. Our study demonstrated significant performance in predicting EGFR and KRAS mutations in NSCLC patients by integrating clinical data with radiomic variables derived from 2,231 medical images, in comparison to the AUC values of the ML-based study and the reference study. Our methodology using CatBoost (AUC = 0.99 ± 0.00) and XGBoost (AUC = 0.99 ± 0.01) for the EGFR mutation scenario, as well as CatBoost (AUC = 0.99 ± 0.00) and XGBoost (AUC = 0.98 ± 0.01) for the KRAS mutations scenario, demonstrates a significant increase in predictive accuracy. In contrast to the 10-fold cross-validation method used in the reference research, we utilized 5-fold cross-validation for a precise evaluation of model performances. The models demonstrated consistency and reliability with a minimum standard deviation. The exceptional performance attained is founded on a cohesive data approach that amalgamates clinical characteristics with extensive radiomic variables, alongside the utilization of sophisticated gradient boosting methods. Upon evaluating the model's performance for EGFR mutation status, four machine learning models were compared, with the CatBoost model demonstrating superior efficacy in predicting the mutation status. The CatBoost model exhibited superior performance with an accuracy of 0.967 ± 0.010, balanced accuracy of 0.931 ± 0.022, precision of 0.968 ± 0.010, sensitivity of 0.991 ± 0.003, F1-Score of 0.979 ± 0.006, and ROC-AUC of 0.989 ± 0.002, compared to the accuracy of 0.888 ± 0.017, balanced accuracy of 0.786 ± 0.038, precision of 0.907 ± 0.017, sensitivity of 0.957 ± 0.006, F1-Score of 0.931 ± 0.010, and ROC-AUC of 0.933 ± 0.021 of another model, as well as the SVM model's accuracy of 0.730 ± 0.015, balanced accuracy of 0.714 ± 0.015, precision of 0.903 ± 0.011, sensitivity of 0.742 ± 0.028, F1-Score of 0.814 ± 0.014, and ROC-AUC of 0.786 ± 0.017. The CatBoost model was succeeded by the XGBoost model, which exhibited comparable performance metrics, demonstrating commendable classification efficacy with an accuracy of 0.963 ± 0.010, precision of 0.969 ± 0.004, and ROC-AUC of 0.985 ± 0.007. The CatBoost and XGBoost models had excellent precision scores of 0.991 and 0.985, respectively. In this instance, false negatives are reduced, and all performance indicators exhibit low standard deviation values, indicating that the models are dependable and consistent. Upon analyzing the model's performance for KRAS mutation status it is evident that the CatBoost model exhibits superior performance across the machine learning measures. The CatBoost model exhibited the following metrics: accuracy (0.965 ± 0.015), balanced accuracy (0.954 ± 0.021), precision (0.953 ± 0.024), sensitivity (0.994 ± 0.007), F1-score (0.973 ± 0.011), and ROC-AUC (0.990 ± 0.005). The accuracy of the Random Forest model was 0.879 ± 0.008 (0.878 - 0.879), balanced accuracy was 0.839 ± 0.010 (0.839 - 0.840), precision was 0.848 ± 0.012 (0.848 - 0.849), precision was 0.986 ± 0.020 (0.985 - 0.987), F1-score was 0.912 ± 0.007 (0.911 - 0.912), and ROC-AUC was 0.942 ± 0.008 (0.942 - 0.942). The performance metrics of the SVM model are as follows: accuracy is 0.813 ± 0.024 (0.812 - 0.814), balanced accuracy is 0.795 ± 0.018 (0.794 - 0.796), precision is 0.846 ± 0.010 (0.845 - 0.846), sensitivity is 0.863 ± 0.043 (0.861 - 0.864), F1-score is 0.854 ± 0.022 (0.853 - 0.855), and ROC-AUC is 0.847 ± 0.013 (0.847 - 0.848). The XGBoost model exhibits an accuracy of 0.951 ± 0.006 (0.950 - 0.951), a precision of 0.935 ± 0.009 (0.934 - 0.935), and a ROC-AUC of 0.982 ± 0.011 (0.981 - 0.982), which closely approximates the performance of the CatBoost method. Integrating biomedical data with data from images has not only improved the prediction accuracy of ML models, but also demonstrated strong generalization capabilities across different data splits.

Conclusion

The performance metrics of the models were evaluated using diverse machine learning methods on biomedical pictures of NSCLC patients and associated clinical data. Model performance indicators are evaluated against the outcomes derived from the integration of biological data and image-related factors. The CatBoost algorithm demonstrated superior classification performance in predicting both EGFR and KRAS mutation statuses.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: E.G., A.P.; Control/Supervision: A.K.A., E.G., A.P.; Analysis and/or Interpretation: E.G., A.K.A., A.P.; Literature Review: A.P., A.K.A., EG; Writing the Article: A.P., E.G; Critical Review: A.P., E.G.; References and Fundings: A.P.; Materials: A.P., E.G Data Avability Statement

This study is a part of the PhD thesis completed by Abdulvahap Pınar in May 2025 and submitted to the Council of Higher Education Thesis Center (https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp). The main text is in Turkish. Only the abstract is available in English.

References

1. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* 48, 441-446 (2012).
2. Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer research* 77, e104-e107 (2017).
3. Chaddad, A., Daniel, P., Sabri, S., Desrosiers, C. & Abdulkarim, B. Integration of radiomic and multi-omic analyses predicts survival of newly diagnosed IDH1 wild-type glioblastoma. *Cancers* 11, 1148 (2019).
4. Song, L. *et al.* Clinical, conventional CT and radiomic feature-based machine learning models for predicting ALK rearrangement status in lung adenocarcinoma patients. *Frontiers in Oncology* 10, 369 (2020).
5. Xu, Y. *et al.* Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research* 25, 3266-3275 (2019).
6. Yamamoto, S. *et al.* ALK molecular phenotype in non–small cell lung cancer: CT radiogenomic characterization. *Radiology* 272, 568-576 (2014).
7. Shi, L. *et al.* Radiomics for response and outcome assessment for non-small cell lung cancer. *Technology in cancer research & treatment* 17, 1533033818782788 (2018).

8. Armanious, K. *et al.* MedGAN: Medical image translation using GANs. *Computerized medical imaging and graphics* 79, 101684 (2020).
9. Razzak, M. I., Naz, S. & Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, 323-350 (2017).
10. Yu, D. *et al.* in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017).* 569-572 (IEEE).
11. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging* 35, 1207-1216 (2016).
12. Little, M. P., Wakeford, R., Tawn, E. J., Bouffler, S. D. & Berrington de Gonzalez, A. Risks associated with low doses and low dose rates of ionizing radiation: why linearity may be (almost) the best we can do. *Radiology* 251, 6-12 (2009).
13. Jacobs, C. *et al.* Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. *European radiology* 26, 2139-2147 (2016).
14. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. Data for NSCLC radiogenomics collection. The Cancer Imaging Archive. 2017;10:K9.
15. Saxena, S. *et al.* Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers* 14, 2860 (2022).
16. Brahmer, J. R. *et al.* The Society for Immunotherapy of Cancer consensus statement on immunotherapy for the treatment of non-small cell lung cancer (NSCLC). *Journal for immunotherapy of cancer* 6, 1-15 (2018).
17. Reck, M. *et al.* Updated analysis of KEYNOTE-024: pembrolizumab versus platinum-based chemotherapy for advanced non–small-cell lung cancer with PD-L1 tumor proportion score of 50% or greater. *Journal of clinical oncology* 37, 537-546 (2019).
18. Armato, S. G. *et al.* Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology* 225, 685-692 (2002).
19. Armato III, S. G., Giger, M. L. & MacMahon, H. Automated detection of lung nodules in CT scans: preliminary results. *Medical physics* 28, 1552-1561 (2001).
20. Wang, C., Elazab, A., Wu, J. & Hu, Q. Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics* 57, 10-18 (2017).
21. Nabiyev, V. Yapay Zeka, Seçkin Yayıncılık San. *Ve Tic. AŞ, Ankara, 724s* (2003).
22. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 15, e1002686 (2018).
23. Russell, S. J. & Norvig, P. (Pearson Education Limited London, UK:, 2016).
24. Karakaya, A. *Meme kanseri tahmininde makine öğrenmesi algoritmaları ve AutoML*, Pamukkale University, (2024).
25. Fan, C., Chen, M., Wang, X., Wang, J. & Huang, B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in energy research* 9, 652801 (2021).
26. Park, K., Chae, M. & Cho, J. H. Image pre-processing method of machine learning for edge detection with image signal processor enhancement. *Micromachines* 12, 73 (2021).
27. Ge, G., Shi, Z., Zhu, Y., Yang, X. & Hao, Y. Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four machine learning algorithms. *Global Ecology and Conservation* 22, e00971 (2020).
28. Breiman, L. Random forests. *Machine learning* 45, 5-32 (2001).
29. Okumus, H. & Nuroglu, F. M. A random forest-based approach for fault location detection in distribution systems. *Electrical Engineering* 103, 257-264 (2021).
30. Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.
31. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
32. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).
33. Yue, S., Li, P. & Hao, P. SVM classification: Its contents and challenges. *Applied Mathematics-A Journal of Chinese Universities* 18, 332-342 (2003).
34. Ibm, C. IBM SPSS statistics for Windows. *Armonk (NY): IBM Corp* (2012).
35. Bakr, S. *et al.* Data for NSCLC radiogenomics collection. *The Cancer Imaging Archive* 10, K9 (2017).
36. Huang, Y. *et al.* Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology* 281, 947-957 (2016).
37. Ball, D. L. *et al.* The complex relationship between lung tumor volume and survival in patients with non-small cell lung cancer treated by definitive radiotherapy: a prospective, observational prognostic factor study of the Trans-Tasman Radiation Oncology Group (TROG 99.05). *Radiotherapy and Oncology* 106, 305-311 (2013).
38. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* 26, 1045-1057 (2013).
39. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Scientific data* 5, 1-9 (2018).