Optimized Lung Cancer Classification with GridSearchCV and Cross Validation Techniques

Hatice Tekiş[1,2*]

[1*]Department of Computer Engineering, Eskisehir Technical University, Tepebaşı, Eskişehir, 26555, Türkiye.
[2]Kapadokya Vocational School, Kapadokya University, Mustafapaşa, Nevşehir, 50420, Türkiye.

1 Introduction

Lung cancer is a disease caused by the uncontrolled growth of abnormal cells in the lungs, which damage the healthy lung cells around them and can spread to other organs in the body [1].

It is one of the most common and deadly cancers in the world, with 1.8 million people dying each year because it is diagnosed at a late stage [2]. Detection of this disease is nowadays possible through medical imaging such as CT, chest x-ray, MRI. However, these medical imaging systems may be inadequate or sometimes limited for detecting cancer at the early stage and may not give reliable results. The development of an innovative computer aided diagnostic (CAD) system is essential for the early detection of the disease and the preparation and management of an appropriate treatment plan. In addition, smoking, passive exposure to tobacco smoke, hereditary predisposition, environmental pollution and upper respiratory tract infections such as tuberculosis play a vital role in the diagnosis and classification of the disease. Considering all of these, the larger the number of parameters, the more complex and more difficult it is to predict and detect the disease with traditional data analysis methods. On the other hand, these complex data pattern can be interpreted using artificial intelligence methods anymore and a diagnostic system can be developed using machine learning or deep learning. In these diagnostic systems, not only lung cancer can be detected, but also the stages of cancer can be classified and determined with complex large data sets. Moreover, the CAD systems are not only diagnostic but also a support system in clinics, helping clinicians in terms of time and labor.

In the literature, two main approaches to the diagnosis and stage classification of lung cancer are more prominent: Image processing (computer vision) based deep learning models and machine learning based analysis methods working on tabular data. These studies mostly used datasets available at the UCI Machine Learning Repository [3] and Kaggle [4].

In [5], an Ensemble Learning based classification method was developed using medical data of patients with and without lung cancer diagnosis. In the study, Random Forest, CatBoost and XGBoosting achieved 98% accuracy, while AdaBoost achieved 96% accuracy. In another study [6] for early detection of lung cancer, classification was performed with XGBoost, Decision Tree, k-Nearest Neigbor (kNN) and Random Forest (RF) machine learning algorithms with a Kaggle dataset [7] containing risk factors such as smoking history, family history, environmental effects; Random Forest is overperformed by the result with 95.16% accuracy. Using the same Kaggle dataset [7] used in the previous study, a hybrid method was proposed by combining Nonlinear regression and Gaussian Mixture Model (NLR-GMM) with this method, classification was performed with 92.88% accuracy and more successful results were obtained than traditional machine learning algorithms Gaussian Naïve Bayes and KNN [8].

Sachdeva, Ravi Kumar, et al. [9] evaluated the classification performance using Support Vector Machine (SVM), Linear Regression (LR), Naive Bayes (NB), RF and kNN on the same dataset; they proposed a new approach, Pearson Correlation Coefficient based Weighted KNN (PCWKNN). This approach outperformed the other algorithms with an accuracy rate of 98.39% and was tested with different disease datasets, emphasizing the generalizability and robustness of the PCWKNN method. Another research study [10] investigated the effect of Genetic Folding Strategy approach on the classification to improve the kernel functions of traditional SVM algorithm, using Kaggle based Lung cancer dataset. The GFS-based model achieved the highest classification success with 96.2% accuracy compared to linear, polynomial and radial basis functions of SVM. In [11], the benign and malignant lung cancer cells were classified using various machine learning algorithms with the lung cancer dataset from the UCI machine learning repository [12]; in the comparative analysis, Radial Basis Function (RBF) indicated the best classification performance with an accuracy of 81.25%. In [13], the effect of various data preprocessing methods on classification performance was investigated using a 32×56 dimensional lung cancer dataset from UCI [12]. The dataset created with nine different preprocessing techniques was evaluated with six different machine learning algorithms; among these preprocessing techniques, the most successful results were obtained with Z-score normalization by 83% accuracy, Principal Component Analysis (PCA) method by 87% accuracy and Information Gain feature selection method by 71% accuracy. In [14] study, synthetic data was generated with conditional tabular generative adversarial networks (CTGAN) and Random Forest algorithm was applied into the synthetic data in order to remove class imbalance and improve the performance of machine learning based classification models on a lung cancer dataset with 309 samples from Kaggle. In the classification performance over nine different machine learning algorithms with balancing methods such as SMOTE, Borderline-SMOTE and SMOTE-ENN, the CTGAN-RF model successfully addressed

the class imbalance problem by achieving 99% precision, recall and F1-score values along with 98.93% accuracy.

In previous studies performed with the Kaggle and UCI datasets, the datasets contain a limited amount of data. While the Kaggle data is 309×16 in size, the dimension of UCI dataset is 32×56. In artificial intelligence studies, the greater the number and variety of data, the greater the generalizability and efficiency of the trained model. In this study, a large number of machine learning algorithms were compared and assessed with the enhanced and updated dataset [15] on Kaggle. A review of the literature indicates that a new study has not been done with this updated dataset. This research is one of the first noteworthy studies. Furthermore, hyperparameter optimization and reliable robustness techniques are not sufficiently used in existing studies. In this research, hyperparameter optimization was performed with GridSearchCV in order to increase the classification success of the models and to improve generalizability in addition to classification with a large number of various machine learning algorithms. Moreover, the overfitting risk of the models was reduced by applying 4-fold cross validation.

2 Methods

2.1 Dataset

In this study, we used a publicly available dataset [15] on Kaggle, for lung cancer diagnosis. The dataset consists of 16 columns (15 feature columns + 1 target variable) and 1,157 patient records. The features in the dataset include patient lifestyle and demographic information such as gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallow of difficulty and chest pain. Target feature contains information about whether the patient has lung cancer (YES/NO). The class distributions are generally balanced. The distribution of classes is visualized in Figure [X] below.

2.2 Data Preprocessing

No missing data was found during the analysis of the dataset. Since the target variable "Lung Cancer" column is categorical in the form of (YES/NO), it was transformed into 0 (NO) and 1 (YES) using label encoding for binary classification. The other categorical columns in the dataset (except AGE and GENDER) were initially coded as YES = 2 and NO = 1, and were relabeled as YES = 1 and NO = 0 to ensure consistency with the target variable.

The AGE column, which is a numeric variable, was normalized by using the StandardScaler method in order for the algorithms to learn better. Figure 1 shows the age distribution of the 1,157 patients in the dataset. Ages of patients vary in the range from 20 to 87 years. The mean age was calculated as 50.75, whereas the median was computed as 54.00. Having analyzed the histogram, it is obviously seen that the majority of the individuals in the dataset are prominent in the 55-65 age range. This shows that lung cancer is more prevalent in middle-aged and older individuals. Additionally, the distribution of the age variable before normalization is far from being symmetric and

demostrates a slightly right-skewed structure. Therefore, an appropriate preprocessing like standardization was applied to the age variable before the model training.

In Figure 2, the left plot illustrates that individuals diagnosed with lung cancer have a higher average age (55.78), whilst healthy individuals have less mean of age (46.73). The right plot indicates the number of individuals in each class.

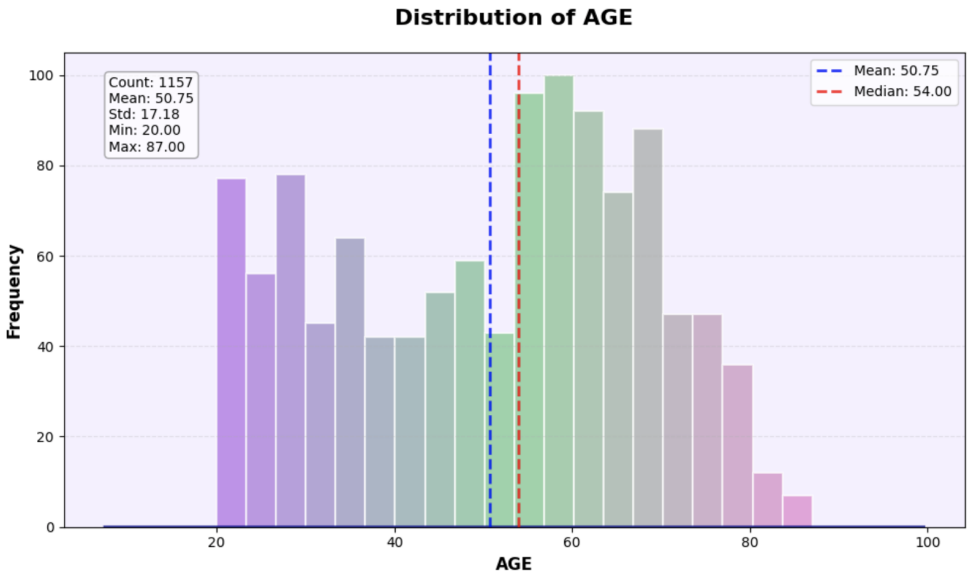

Fig. 1 Distribution of AGE column

The GENDER column, which contains gender information, was labelled as M = 0 and F = 1 since it is in string format (M/F).

Figure 3 shows the Pearson correlation coefficients between the features in the dataset. It clearly indicates that the variables "chest pain", "swallowing difficulty", "smoking" and "shortness of breath" have a high positive correlation with the target variable "lung cancer". Individuals with these symptoms have a strong likelihood of having lung cancer. On the contrary, the correlation with the variables "gender", "allergy", and "alcohol consuming" is quite low, indicating that not much information can be obtained from these variables in model training.

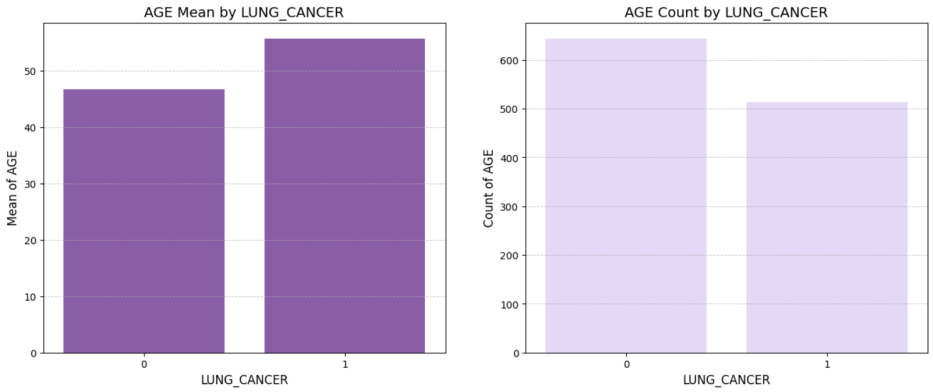

Fig. 2 The numerical distribution of the AGE featuure by LUNG CANCER target variable in terms of mean and count.

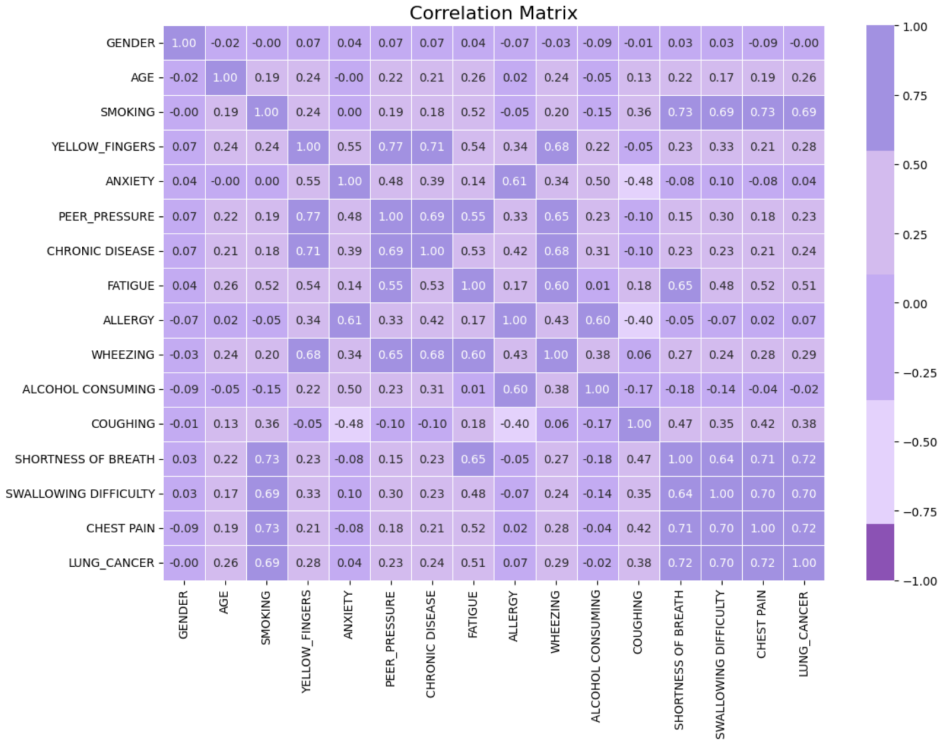

Fig. 3 Correlation matrix of the dataset

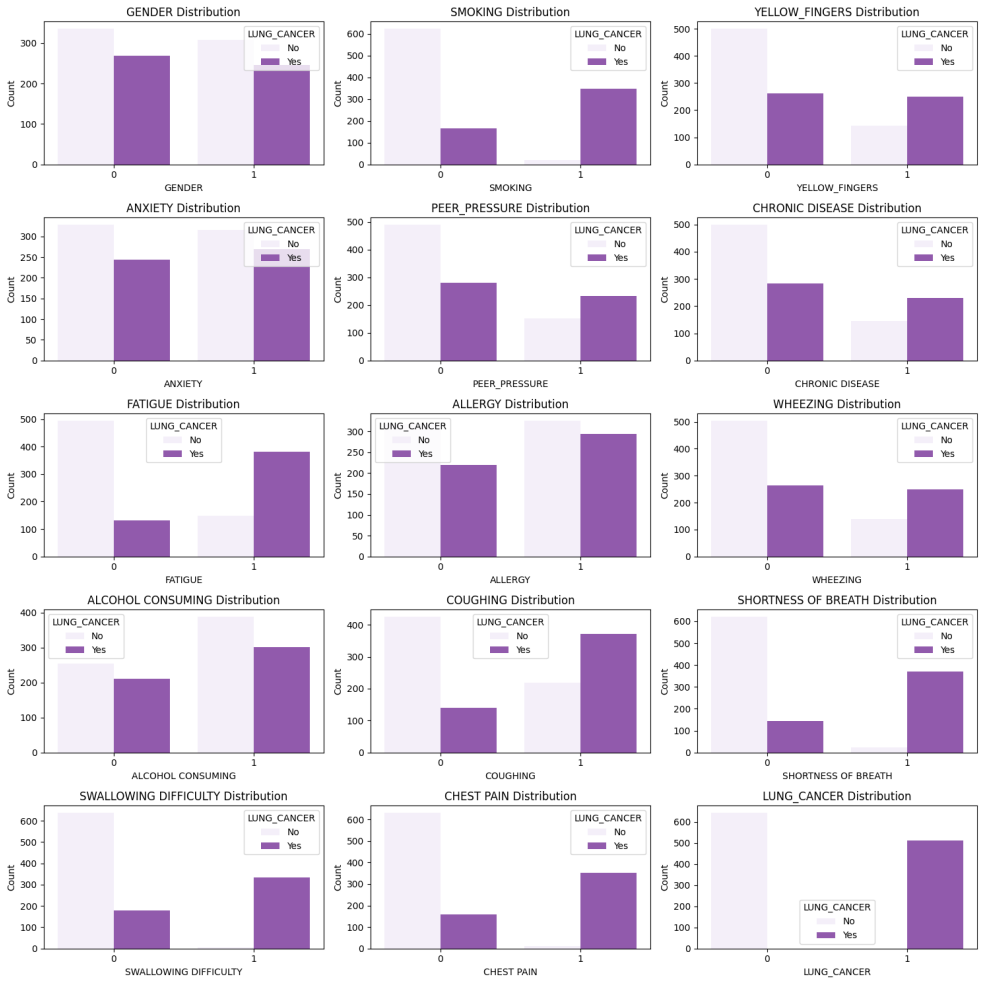

Fig. 4 Distribution of categorical features according to LUNG CANCER variable.

2.3 Model Training and Hyperparameter Optimisation

The data set was divided into 80% train and 20% test. The training data (80% part) was trained with 8 different machine learning algorithms by applying 3 repetitions of 4-fold stratified cross-validation (Repeated Stratified K-Fold Cross-Validation). By applyinf cross calidation technique, we aimed to reduce the possibility of overfitting over the models. The ML algorithms used are Random Forest, SVM, KNN, XGBoost, LightGBM, GradientBoosting, AdaBoost, CART. In addition, hyperparameter optimization was performed using GridSearchCV to improve the performance of the model. The hyperparameters and best parameters are shown in Table 1 below. The trained

models were saved and tested with test data. The performance of the models was measured by accuracy, precision, recall, F1-score metrics and the classification abilities of the models were compared.

Model	Hyperparameter Search Space

SVM	`C: [0.1, 1, 10]`
	`kernel: ['rbf', 'linear']`
	`gamma: ['scale', 'auto', 0.1]`

Gradient Boosting	`n_estimators: [100, 200, 300]`
	`learning_rate: [0.001, 0.01, 0.1]`
	`max_depth: [3, 5, 8]`
	`min_samples_split: [10, 20]`

AdaBoost	`n_estimators: [50, 100, 200]` `learning_rate: [0.01, 0.1, 1]`

Random Forest	`max_depth: [8, 15]`
	`min_samples_split: [15, 20]`
	`n_estimators: [200, 300]`

XGBoost	`learning_rate: [0.01, 0.001, 0.0001, 0.00001]`
	`max_depth: [5, 8, 10]`
	`n_estimators: [100, 200, 300, 400, 500]`
	`colsample_bytree: [0.8, 1]`
	`subsample: [0.8, 1]`
	`gamma: [0, 1]`
	`min_child_weight: [1, 3]`

LightGBM	`learning_rate: [0.01, 0.001, 0.0001, 0.00001]` `n_estimators: [100, 200, 300, 400, 500]`
KNN	`n_neighbors: range(2, 50)`
CART	`max_depth: range(1, 20)` `min_samples_split: range(2, 30)`

Table 1 Hyperparameter search spaces for model optimization.

3 Results

We compared the performance of eight distinc machine learning algorithms using the metrics of accuracy, precision, recall and F1-score. In-depth classification results on the test dataset are illustrated in Table 4.

In our analysis, the Support Vector Machine (SVM) outperformed the best with an accuracy of 95.3%, an F1-score of 94.7%, an Precision of 98%, and an Recall of 91.6%. These results indicate that SVM achieves a great balance of correctly detecting positive cases and decreasing the number of misclassifications. The Random Forest algorithm performed second best algorithm for classfying the lung cancer, achieving results very similar to the SVM in all evaluation metrics.

Boosting-based models (LightGBM, AdaBoost, Gradient Boosting and XGBoost) presented remarkable classification results. In particular, the high precision values of these algorithms show that they minimize the number of false positives. Contrarily,

K-Nearest Neighbors (KNN) and Decision Tree (CART) underperformed for classification, despite of acceptable results, when compared to ensemble learning and kernel-based algorithms.

The results of this study obviously show the higher effectiveness and performance of SVM and ensemble learning approaches, especially when applied repeated robust cross-validation techniques and hyperparameter optimization.

Model	Accuracy	Precision	Recall	F1 Score	AUC
XGBoost	0.9150	0.9452	0.8572	0.8984	0.9493
LightGBM	0.9168	0.9453	0.8613	0.9008	0.9523
SVM	0.9261	0.9565	0.8720	0.9119	0.9580
GradientBoosting	0.9232	0.9564	0.8654	0.9082	0.9564
AdaBoost	0.9121	0.9548	0.8400	0.8934	0.9348
KNN	0.9175	0.9603	0.8474	0.9001	0.9337
CART	0.9117	0.9448	0.8491	0.8938	0.8891
Random Forest	0.9186	0.9423	0.8687	0.9035	0.9403

Table 2 Performance metrics of base models using train data.

Model (Tuned)	Accuracy	Precision	Recall	F1 Score	Best Params

XGBoost	0.9132	0.9524	0.8449	0.8951	`colsample_bytree=0.8`
					`gamma=0`
					`lr=0.01`
					`max_depth=5`
					`min_child_weight=1`
					`n_est=100`
					`subsample=1`

LightGBM	0.9178	0.9499	0.8588	0.9016	`lr=0.01` `n_est=400`

SVM	0.9261	0.9565	0.8720	0.9119	`C=1`
					`gamma=scale`
					`kernel=rbf`

GradientBoosting	0.9200	0.9541	0.8597	0.9041	`lr=0.01`
					`max_depth=5`
					`min_samples_split=20`
					`n_est=200`

AdaBoost	0.9114	0.9529	0.8399	0.8926	`lr=0.1` `n_est=200`
KNN	0.8998	0.9674	0.7989	0.8749	`n_neighbors=44`
CART	0.9168	0.9367	0.8703	0.9018	`max_depth=8` `min_samples_split=23`

Random Forest	0.9193	0.9355	0.8777	0.9052	`max_depth=8`
					`min_samples_split=15`
					`n_est=200`

Table 3 Performance metrics and best hyperparameters of tuned models.

Model	Accuracy	Precision	Recall	F1 Score
SVM	**0.953**	0.980	0.916	**0.947**
Random Forest	0.948	0.970	0.916	0.942
LightGBM	0.940	0.979	0.888	0.931
AdaBoost	0.935	0.979	0.879	0.926
Gradient Boosting	0.935	0.979	0.879	0.926
XGBoost	0.931	0.979	0.869	0.921
CART	0.927	0.959	0.879	0.917
KNN	0.922	0.978	0.850	0.910

Table 4 Comparison of test performance metrics across eight machine learning models.

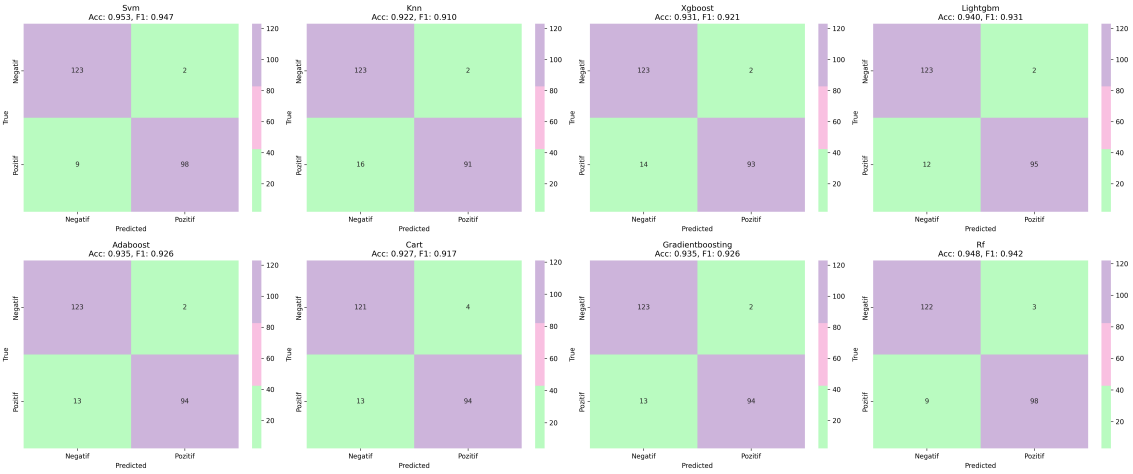

Fig. 5 Confusion matrices of eight different machine learning models

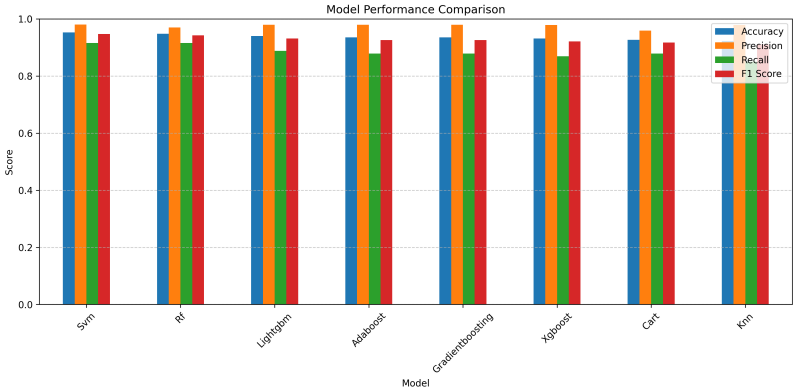

Fig. 6 Model performance comparison

References

[1] Saha, P., Nyarko, R.O., Lokare, P., Kahwa, I., Boateng, P.O., Asum, C., *et al.*: Effect of covid-19 in management of lung cancer disease: A review. Asian Journal of Pharmaceutical Research and Development **10**(3), 58–64 (2022)

[2] Choudhary, D.M., Roshan, V., Sri, A.D., Ajith, J., Srinivas, P., *et al.*: Implementation of predictive modeling for lung cancer diagnosis using machine learning and deep learning algorithms on lung dataset. In: 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), pp. 960–966 (2024). IEEE

[3] UCI Machine Learning Repository. https://archive.ics.uci.edu/datasets. Accessed: 2025-05-13

[4] Kaggle. https://www.kaggle.com. Accessed: 2025-05-13

[5] Yanuar, R., Sa'adah, S., Yunanto, P.E.: Implementation of hyperparameters to the ensemble learning method for lung cancer classification. Building of Informatics, Technology and Science (BITS) **5**(2), 498–508 (2023)

[6] Raigonda, M.R., Mama, G., Bainoor, R.: Lung cancer detection using machine learning techniques

[7] Lung Cancer Dataset on Kaggle. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Accessed: 2025-05-13

[8] Rajaguru, H., SR, S.C., Chidambaram, S.: Gaussian mixture model based hybrid machine learning for lung cancer classification using symptoms. In: 2022 Smart Technologies, Communication and Robotics (STCR), pp. 1–4 (2022). IEEE

[9] Sachdeva, R.K., Bathla, P., Rani, P., Lamba, R., Ghantasala, G.P., Nassar, I.F.: A novel k-nearest neighbor classifier for lung cancer disease diagnosis. Neural Computing and Applications **36**(35), 22403–22416 (2024)

[10] Mezher, M.A., Altamimi, A., Altamimi, R.: A genetic folding strategy based support vector machine to optimize lung cancer classification. Frontiers in Artificial Intelligence **5**, 826374 (2022)

[11] Patra, R.: Prediction of lung cancer using machine learning classifier. In: Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1, pp. 132–142 (2020). Springer

[12] Hong, Z.Q., Yang, J.Y.: Lung Cancer. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C57596 (1991)

[13] Gültepe, Y.: Performance of lung cancer prediction methods using different

classification algorithms. Computers, Materials & Continua **67**(2) (2021)

[14] Alzahrani, A.: Early detection of lung cancer using predictive modeling incorporating ctgan features and tree-based learning. IEEE Access (2025)

[15] more-accurate-lung-cancer-dataset on Kaggle. https://www.kaggle.com/datasets/chandanmsr/more-accurate-lung-cancer-dataset. Accessed: 2025-05-13