**Lung Cancer Detection with Machine Learning Supported Image Processing Techniques**

***Furkan ÜSTÜNER1 [C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0000-1451-6866), Mustafa SAZ2[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0001-1150-4568), İdris GENÇ3[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0009-9392-3878)Mahmut BÜYÜKBAŞ4,[[1]](#footnote-1)\*[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0003-4106-7749)***

*1,2,3,4 Faculty of Engineering, Electrical and Electronics Engineering, Abdullah Gül University, Kayseri, Turkey*

|  |
| --- |
| **Abstract**  Fast processes play an important role in the diagnosis of lung cancer. This study aims to develop a computer-aided diagnostic (CAD) system using machine learning algorithms and advanced image processing techniques. The dataset used contains computed tomography (CT) scans obtained from two different private hospitals in Iraq and considers healthy individuals as well as lung cancer patients at different stages of the disease. The process has three important stages. The first is initial image preprocessing to get better output and improve image quality, followed by segmentation and feature extraction to identify relevant features, show the diseased area, and finally feature selection to optimize the inputs and make the best choices for the classification stage . In the project, various machine learning algorithms such as random forest, decision trees and neural networks are tested to distinguish benign and malignant cases, and the most ideal classification method for the data set is selected. The performance of these classification methods is evaluated using metrics such as accuracy, precision, and F1 score to ensure the reliability of the system. This study aims to significantly increase the effective treatment of patients by contributing positively to the lung cancer diagnosis process. |
| Keywords: Lung Cancer Detection, Image Processing, CT Scan Analysis, Classification, Feature Extraction |

1. **Introduction**

Lung cancer is one of the most lethal types of cancer and millions of people die from this disease every year. Accurate diagnosis of lung cancer plays a critical role in the success of treatment. Computer-aided diagnosis (CAD) systems provide visualization of medical data using advanced technology and can significantly speed up the diagnosis process. In our project, computed tomography (CT) images of healthy and lung cancer patients at different stages in the IQ-OTH/NCCD dataset obtained from two private hospitals in Iraq were used[1]. Our aim is to accurately classify tumors as healthy, benign and malignant using image processing and machine learning algorithms. We aim to increase the classification accuracy, sensitivity and F1 score. In this direction, methods such as Contrast, Energy, Gabor filters and HomogeneityGLCM are used in the segmentation and feature extraction stages. Hasan Hejbari Zargar and his team demonstrated the effectiveness of these techniques by reaching 92.08% sensitivity, 91% accuracy and 93% AUC values ​​with the VGG16 algorithm [2]. For classification and performance evaluation, algorithms such as artificial neural networks, decision trees and random forests are tested, and the correct classification success of each algorithm is examined. This model will provide significant contributions to correct diagnosis and appropriate treatment planning with high accuracy, sensitivity and F1 score.

1. **Materials and Methods** 
   1. **Importing of the Datasets**

Importing datasets is used in the first phase of a machine learning project and is especially used when it comes to diagnosis using computer-aided imaging. This study divided the dataset into three different cases: benign, malignant, and normal. After the necessary definitions were made using the Python coding language, the files were downloaded to the computer. This setting makes it easier to manipulate data programmatically. It helps the data to be used in subsequent stages such as pre-processing, feature extraction, feature selection and classification. Keeping the data set organized is an important way to ensure the success of the project and plays an important role in increasing efficiency in the analysis process. Additionally, a sample image from the dataset is included as shown in Figure 1.

siyah beyaz, metin, taslak, siyah içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Importation of the dataset

* 1. **Image Processing Procedures**

In medical imaging, image processing methods are of great importance in the detection of diseases. Processes such as median filter, edge detection and Gaussian blur have been used to minimize noise, identify morphological changes and select important features. Color images are reduced to a single channel by converting to grayscale, thus reducing data size and increasing processing efficiency. Gaussian filter provides clearer and more processable images by reducing noise in sensitive medical data. Median filter provides clearer tumor segmentation in CT scans by eliminating random noise (e.g. salt and pepper). Edge detection provides detailed analysis by emphasizing density changes in the image and highlighting tumor boundaries. These methods contribute to a clearer view of anatomical structures in particular.You can also see the image processing operations performed on benign cases (Figure 2), malignant cases (Figure 3), and normal cases (Figure 4).

metin, taslak, siyah, siyah beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, daire, siyah beyaz, taslak içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Benign Cases Figure . Melignant Cases

metin, taslak, siyah beyaz, siyah içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Normal Cases

* 1. **Segmentation**

The operations performed in the previous steps make the data obtained in the segmentation section clearer and more accurate. Particularly in medical imaging and in our project, segmentation separates diseased tissues from healthy tissues, making them more visible and increasing the accuracy of diagnoses at the stage of medical diagnosis. Segmentation on computed tomography (CT) scans used to diagnose lung cancer is crucial for determining the precise boundaries of tumors or other abnormal structures. This study used the segmentation method to mark and examine malignant (Figure 6) and benign (Figure 5) structures within the lung tissue. Segmentation on computed tomography (CT) scans used to diagnose lung cancer is crucial for determining the precise boundaries of tumors or other abnormal structures. This study used the segmentation method to mark and examine malignant (Figure 6) and benign (Figure 5) structures within the lung tissue. Clearer and more accurate data was obtained during feature extraction due to the separation of two different tissue parts. As a result, this separation process helps the machine learning procedures we use for classification work more effectively.

ekran görüntüsü, siyah, kare, kalıp, desen, düzen içeren bir resim

Açıklama otomatik olarak oluşturuldu ekran görüntüsü, kare, kalıp, desen, düzen, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Benign Cases Segmentation Figure . Melignant Cases Segmentation

* 1. **Feature Extraction and Feature Selection**

Feature extraction is an important step in machine learning and data analysis aimed at obtaining meaningful and more processable information from the raw data set. It is also the process of extracting the necessary information from the raw data set, aiming to facilitate the classification prediction of the model. The features extracted from the data aim to increase the accuracy of the model and its generalization capacity in different situations. Feature extraction also reduces the size of the raw data set to make the data more manageable due to the large size of the raw data set. Thus, while the data set is represented in a more concise form, the processing time is reduced, and performance is increased due to the elimination of unnecessary information. In our project, we tried to perform feature extraction of Computerized Tomography (CT) images, which is a critical step in detecting and analyzing pathological conditions of lung CT images. In addition, the 50 different features extracted in this project allow us to analyze different aspects and textural qualities of the images in detail. ANOVA feature selection method was used to identify the extracted features associated with tumourous tissues. ANOVA analyses group means and variances. ANOVA feature selection helps to determine which features are relevant to tumourous tissue. ANOVA calculates the differences between the means of different groups, and if the calculated F value is within the desired range (p<0.05) Figure 7 and Figure 8, it indicates that the feature makes a statistically significant difference between the groups.

metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu ekran görüntüsü, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . ANOVA F-Value for Different Features Figure . ANOVA p-Value for Different Features

The feature being important for the model and having a strong relationship with the target helps to create a model with a more understandable and general structure by increasing the model performance. Selection of meaningful features reduces the risk of overfitting and allows elimination of misleading data. The features that are important in our project are: Correlation expresses pixel relationship; HomogeneityGLCM is effective in showing textural consistency and separating tumor tissue; Elongation determines tumor shape differences; Sphericity separates tumor tissue from regular structures; Eccentricity helps to understand the morphological structure of the tumor; Flatness contributes to the analysis of tissue abnormalities; Entropy (GLSZM) and Roundness are used to detect irregular tumor shapes.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Top 8 Features Based on ANOVA F-Value Figure . Top 8 Features Based on ANOVA p-Value

ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturulduöykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Histogram Graph of Selected Features

In summary, meaningful and important features in an image are the data processed for tumor detection. These features are essential for disease classification and detection in the later stages and incorrect feature extraction and selection of these features is directly related to the model's performance. Feature extraction and selection, which has a direct impact on the training and performance of the model, is essential for the other stages of the project.

1. **Results and Discussion** 
   1. **Classification**

In machine learning, classification is the process of dividing a dataset into subcategories. Especially thanks to the classification techniques used in the field of medical imaging, the images obtained are divided into groups in terms of identifying diseases. In this project, classification methods such as decision trees, Deep Neural Networks (DNN) and Random Forests were used (Figure 12). The Random Forest method is an interesting approach; it brings together many decision trees and evaluates the results produced by each tree and selects the most common result. Another advantage of this method is that it provides diversity and a wide perspective, because each decision tree looks at the data differently. The advantage of this method is that it provides high accuracy rates and reduces the risk of overfitting. Deep Neural Networks (DNN), which are high-layer artificial neural networks, are also a very successful classification method in recognizing and classifying complex patterns. (Figure 13). This allows the DNN to discover small details in the images. Decision trees are a simple but effective method that divides the data according to certain criteria and classifies the results by creating decision points on each branch (Figure 14). The feature selection research in the project highlighted that these three classification methods are suitable for lung cancer diagnosis. We ranked the features extracted from our dataset according to their importance and selected the most significant ones. This strategy greatly increased the success and accuracy of our model. In particular, selecting the right features can be life-saving for patients about to be diagnosed. This method aims to obtain more robust and reliable results by examining.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Decision Trees Clasification Figure . DNN (Deep Neural Network) Classification

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Random Forest Classification

Three different classification models (Random Forest, Decision Trees and Deep Neural Network, also known as DNN) were measured and their performances were compared. Each of the models was evaluated using important metrics such as accuracy, recall, precision and F1 score. Observations and graphical result (Figure 15) show that the DNN model is better than other classification models in every aspect. For this reason, we decided to use the Deep Neural Network (DNN) method because it provides better results and is better suited to the data set. This decision was made after examining the models in more depth.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Model Comparison Table

* 1. **Cross Validation**

Cross-validation is used to evaluate the accuracy of classification models and avoid overfitting. Particular attention should be paid to complex structures such as Deep Neural Networks (DNN). Cross-validation is based on the principle of splitting a dataset into random subsets and using each subset as testing and training sets. This method is known as “k-fold cross validation”; where "k" indicates how many subsets the "k" dataset will be divided into. This increases the reliability of the model. The DNN model used in the project uses cross-validation to accurately evaluate how effective selected features are in diagnosing lung cancer. This prevents the model from fitting too many subsets of data. Once feature selection is completed in the project, evaluating the impact of the DNN model on various datasets allows determining the best feature combinations. This procedure is important to understand how the model will respond to various patient data. Cross-validation improves the final diagnostic success of the project by increasing the reliability of the model prior to medical application. In the project, various studies were carried out to increase the number of layers in the structure of the Deep Neural Networks (DNN) model in order to increase its effectiveness. Biden was tested by trying multilayer numbers. DNNs leverage multiple hidden layers to gain insight into complex data views. These layers affect the accuracy of predictions because each processes data differently. Therefore, the number of layers in the DNN model had to be adjusted to determine the most suitable structure for the project. As a result of our observations, increasing the number of layers allows the model to understand the data better, but increasing the number of layers too much negatively affects the generalization ability of the model. Therefore, experiments were conducted to determine the ideal number of layers that meet the highest accuracy and generalization performance in the diagnosis of lung cancer images in our dataset.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Effect of change in the number of layers

The work done in the project to maximize the performance of the model using Deep Neural Networks (DNN) changed not only the number of layers but also the number of epochs and test set size. There is an epoch number, which indicates how many times the entire training data set will be processed in the learning process. Increasing the number of epochs allows the model to learn the training data better, but too many epochs may cause the model's generalization ability to decrease and overfit. The project tried to find the ideal value that tried to maximize the model's performance on both training and validation datasets by trying different epoch numbers.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Effect of change in the number of Epoch Model Accuracy

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Effect of change in the number of Epoch Model Loss

However, adjustments to the test set size are important to the accuracy and reliability of the model. The test set allows the model to be compared with data it has not seen during the training process. This is usually set as a certain percentage of the data set. As the size of the test set increases, more accurate predictions can be made about how the model will perform with real data. However, a test set that is too large may be negatively affected as it will reduce the amount of data required for training the model. The project efforts achieved an ideal balance between the model's training dataset and validation dataset by manipulating the test set size, optimizing the success of the model both in the training process and in real-world conditions. These changes helped the DNN model provide high accuracy and reliability in diagnosing lung cancer.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure . Effect of change in the ratio of Test Size

1. **Conclusion**

This study examines a new CAD system for lung cancer diagnosis. This project aims to significantly increase patient survival rates through rapid diagnosis, supported by data collected from private hospitals in Iraq. This study distinguishes benign and malignant cases by detailing processes such as image preprocessing, segmentation, feature extraction and feature selection. Various machine learning algorithms are used, such as SVMs, decision trees, and artificial neural networks. The findings support our system's correct detection capability, supported by high accuracy, sensitivity, and metrics such as F1 score. As a result, it seems that this project contributes to the development of treatment methods by creating an effective tool in the rapid diagnosis of lung cancer and has the potential to extend the lifespan of patients. These achievements show once again what impact technological innovations will have on health and how important rapid diagnosis is.

**References**

1. Mahimkar, A. (2021). IQ-OTH/NCCD lung cancer dataset [Data set]. Kaggle.
2. Zargar, H. H., Zargar, S. H., Mehri, R., & Tajidini, F. (2023). Using VGG16 Algorithms for classification of lung cancer in CT scans Image. arXiv preprint arXiv:2305.18367.
3. Alsaadi, E. M. T. A., & Rahman, Z. H. A. A. (2023, December). Automatic lung cancer recognition in chest CT-scan images using SVM classifier. In AIP Conference Proceedings (Vol. 2977, No. 1). AIP Publishing.

1. \* Corresponding author. *e-mail address: mahmut.buyukbas@agu.edu.tr* [↑](#footnote-ref-1)