**Classification of X-ray Images of Atelectasis and Pneumonia**

***Mücahit Ensar ÖZTÜRK1***[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0007-8515-8120)***, Ebubekir BABAYİĞİT1***[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0000-3542-7651)***, Mahmut BÜYÜKBAŞ1***[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0003-4106-7749)***,***

***Bekir Hakan AKSEBZECİ2,***[[1]](#footnote-2)[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0001-7476-8141)

*1Faculty of Engineering, Electrical-Electronics Engineering, Abdullah Gül University, Kayseri, Turkiye*

*1Faculty of Engineering, Electrical-Electronics Engineering, Abdullah Gül University, Kayseri, Turkiye*

*1Faculty of Engineering, Electrical-Electronics Engineering, Abdullah Gül University, Kayseri, Turkiye*

*2Faculty of Engineering, Biomedical Engineering, Abdullah Gül University, Kayseri, Turkiye*

|  |
| --- |
| **Abstract**  Atelectasis and pneumonia are serious lung diseases that can lead to serious complications, including death. The gold standard for diagnosing these diseases is X-ray imaging. This study focuses on the classification of atelectasis and pneumonia diseases through chest X-ray images. A dataset that includes 744 X-ray images was used. In this work, different image processing, feature extraction, feature selection, and classification techniques are examined to improve the performance of medical image classification. In the image processing step, resizing, Gaussian filtering and adaptive histogram equalization methods were applied to the images. Then, region of interests were detected in the segmentation step. In the feature extraction step, first-order statistical features, texture-based features, morphological features and shape-based features were extracted. Information gain, wrapper and correlation-based feature selection methods were applied in the feature selection step. In the classification part, five different machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines, K-Nearest Neighbor and Decision Tree were utilized to classify X-ray images. The results of this study demonstrate that selecting the most effective feature selection and classification methods significantly improves accuracy rates. |
| Keywords: Chest X-ray images, Atelectasis, Pneumonia, Image processing, Machine learning |

1. **Introduction**

A collapsed and non-aerated portion of the lung parenchyma is referred to as atelectasis. This condition occurs when the air sacs in that area close or deflate due to the suffocation of a certain part of the lungs. Acute lung parenchymal infection caused by a variety of microorganisms is known as pneumonia [1, 2]. Pneumonia is a disease caused by infection and inflammation in the lungs. It usually occurs when bacteria, viruses, fungi, or parasites settle in the lungs. Pneumonia is characterized by inflammation of the lung tissue and infection of the alveoli (air sacs). This condition can affect respiratory function and seriously affect respiratory system health. In addition, the symptoms of both diseases can be similar. Common symptoms include shortness of breath, cough, sputum production, chest pain, fever, fatigue, and weakness. In the diagnosis of both diseases, the signs and findings on X-ray images are of great importance. However, manually reviewing these images can be time-consuming and can lead to misdiagnosis. At this point, the use of image processing and machine learning techniques plays a crucial role.

Many studies have been conducted to identify atelectasis and pneumonia in recent years. For instance, on a study that made with 200 images for dataset, samples with atelectasis were classified correctly at a rate of 82% using Convolutional Neural Network (CNN) [3]. On another study that includes 5840 images for dataset, researchers reached 84.1% accuracy with Neural Network model in identifying pneumonia [4]. Wang et al. [5] conducted a comprehensive quantitative performance benchmarking study on eight common thoracic pathology classifications and weakly-supervised localizations using the ChestX-ray8 database. They achieved an AUC (Area-Under-Curve) of 0.7069 for atelectasis and 0.6333 for pneumonia classification with different Deep Convolutional Neural Network (DCNN) models.

In this article, the focus is on the use of X-ray images for the diagnosis of lung diseases that are atelectasis and pneumonia and the importance of image processing and machine learning techniques in classification of these images. The reason of the selecting atelectasis and pneumonia from “ChestX-ray14” dataset is their similarity on X-ray images as can be seen on Figure 1(a) and (b). They cannot easily determined even by medical experts. It can be easier to distinguish these two diseases from each other with image processing and machine learning.

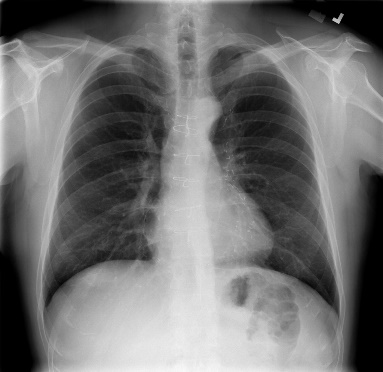
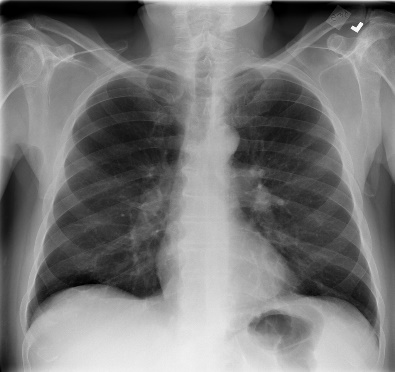
 (a) (b)

Figure 1. X-ray images: (a) atelactasis, (b) pneumonia.

1. **Materials and Methods**

Image processing techniques play an important role in the diagnosis of lung diseases such as atelectasis and pneumonia. These techniques allow the identification of diseased regions, the extraction of features, and use for classification through the processing and analysis of X-ray images. This paper focuses on extracting features from X-ray images by applying machine learning techniques and diagnosing disease using classification algorithms. Feature extraction is used to transform raw data into numerical features. After that, feature selection is applied to find best representative features. Finally, classification algorithms are used to classify diseases that are atelectasis and pneumonia.

## 2.1. Dataset

The study utilized the 'ChestX-ray14 dataset, which includes 14 classes [5, 6]. However, only the atelectasis and pneumonia classes were utilized in this study due to the presence of other diseases and materials (such as pacemakers) in some images. Table 1 presents the distribution of the X-ray image dataset for the two disease classes used in this study.

**Table 1.** X-ray image dataset distribution by diseases.

|  |  |
| --- | --- |
| **Disease Type** | **Number of Images** |
| Atelectasis | 372 |
| Pneumonia | 372 |
| **Total** | **744** |

## 2.2. Image Processing

Image processing steps are important to improve the quality of the X-ray images obtained and to perform more effective analysis. After the dataset selection, image processing steps were applied in order to perform feature extraction efficiently. These steps in below were applied separately and shown in Figure 2.

1. Resizing of the X-ray images
2. Adaptive histogram equalization (AHE)
3. Contrast limited adaptive histogram equalization (CLAHE)
4. The Gaussian filtering after AHE
5. Bilateral filtering after CLAHE
6. Harmonic filtering

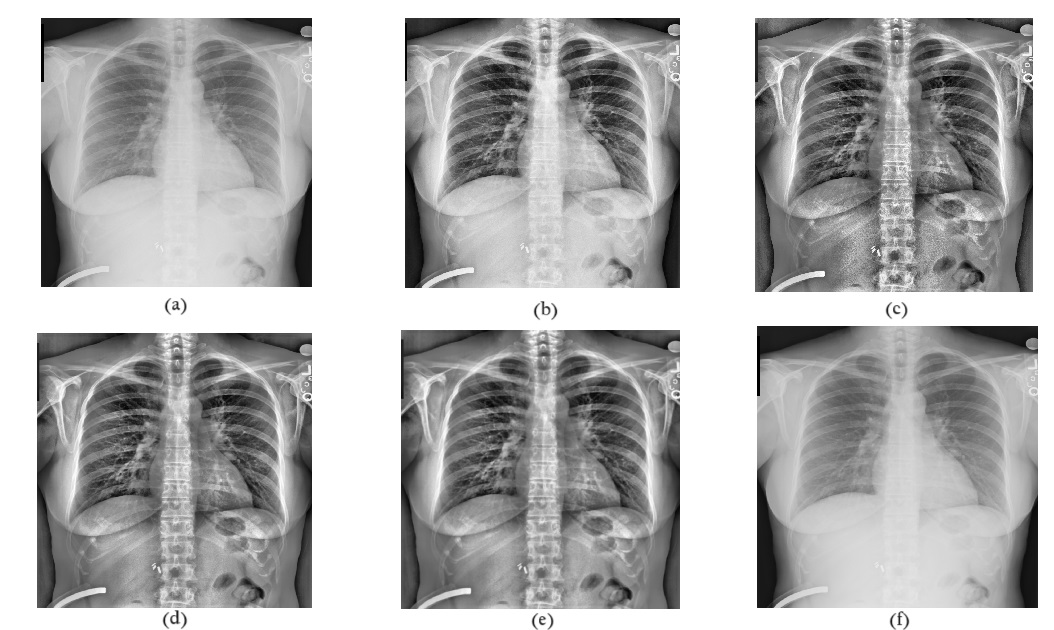
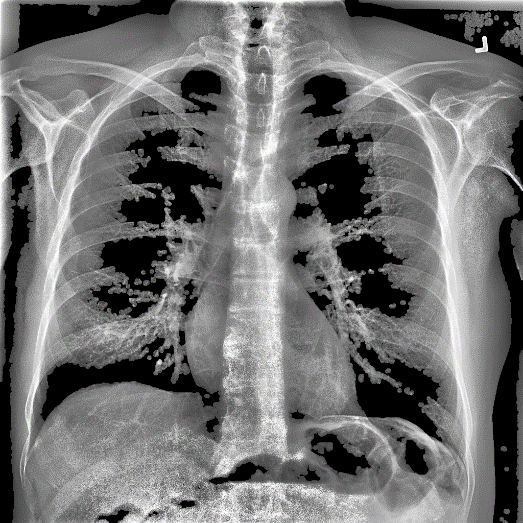


Figure 2. Image processing steps: a) original image of atelectasis, b) after AHE, c) after CLAHE, d) Gaussian filtering after AHE, e) Bilateral filtering after CLAHE, f) after harmonic filtering.

## 2.3. Segmentation

Segmentation is a significant step for detecting the region of interest (ROI) in images. Two different methods were used for the segmentation part. These methods are thresholding segmentation and edge-based segmentation. A certain threshold value is determined for the thresholding segmentation part. Then, the images were converted into binary masks according to this threshold value. After that, the ROIs were found by combining the resulting binary masks and original images. At the same time, the candy edge method, which is one of the edge-based segmentation methods, was applied. Edges within the image were detected and then ROIs were detected by *k*-means clustering. A sample ROI image is shown in Figure 3.

Figure 3. ROI image

## 2.4. Feature Extraction

According to the different features extracted from the ROIs, various information was obtained from its intensities. Extracted features and their numbers are given in Table 2. These features are given in a list below:

* First-order statistical features (Mean, median, standard deviation, minimum and maximum values, and more)
* Texture-based features (GLCM and GLSZM)
* Morphological features (Area, perimeter, compactness, stiffness, and more)
* Shape-based features (Eccentricity, Major Axis Length, Minor Axis Length, Orientation, and more)

**Table 2.** Distribution of the extracted features.

|  |  |
| --- | --- |
| **Features** | **Number of Features** |
| First-order statistical | 34 |
| Texture-based features | 150 |
| Morphological | 20 |
| Shape-based | 30 |
| **Total** | **234** |

## 2.5. Feature Selection

MATLAB has been used in the studies carried out up to this stage [7]. Three distinct feature selection strategies were chosen over Weka [8]. First, the "CfsSubsetEval" (CFS) was chosen, which is based on the correlation-based feature selection technique. Using correlation criteria, this technique finds the sequence of features by their order of importance in a dataset. Then, the information-based feature selection algorithm, "InfoGainAttributeEval" (InfoGain) was chosen. The InfoGain method seeks to choose attributes that provide a high level of information gain. InfoGain computes information gain by calculating the association of each attribute to the target class. Finally, the "WrapperSubsetEval" (Wrapper) was chosen, which is based on the learner-based feature selection technique. The performance of the classification model is used to evaluate the significance of features in this strategy.

## 2.6. Classification

After the feature selection part, following classification algorithms are applied. Naive Bayes computes the class probabilities of the samples. Using Bayes' theorem calculates the chance that a sample belongs to a certain class and assigns the sample to the class with the highest probability. Logistic Regression evaluates the probability that the data belong to a specified class. The model is tailored to the dataset and optimal weight values are obtained during the training phase. Support Vector Machine (SVM) strives for maximal marginal separation. In other words, it seeks the greatest disparity between classes. Kernel was selected as radial basis function for SVM classifiers in this study. K-Nearest Neighbor (KNN) is based on a measure of sample similarity. For KNN classifiers *k* = 5 was chosen in this study. Decision Tree attempts to select the best split. Splits are evaluated using metrics such as information gain, the Gini index, and the complexity criterion. The classification algorithms were applied to the dataset using the 10-fold cross validation method.

1. **Results and Discussion**

In this study, different image processing, feature extraction, feature selection, and classification methods are examined to improve the performance of X-ray image classification. The results in classification methods were compared according to three feature selection methods such as InfoGain, CFS and Wrapper. First of all, InfoGain method was applied for different numbers of feature selection. The results were given in Table 3. It is seen that the classification success increases when the number of features is increased. Secondly, the CFS method was tested by selecting different feature numbers. The results were given in Table 4. Choosing different numbers of features did not cause a change for the CFS method. Finally, the Wrapper method was tested on different classification methods and all results were compared. Since the Wrapper method selects the features that give the best results, there is no need to try different numbers of feature selection.

The results obtained using different feature selection methods and classification methods are shown in Table 5. According to these results, the accuracy rates of different combinations differ from each other. For the Naive Bayes classification method, the highest accuracy rate was 81.04% using the Wrapper method. In the Logistic Regression method, the highest accuracy rate was obtained by using the Wrapper method with 83.87%. The highest accuracy rate for the SVM classification method was obtained at 84.40% with all features. The highest accuracy rate for the KNN method was obtained using the CFS method at 74.32%. In the Decision Tree classification method, the highest accuracy rate of 80.10% was obtained by using the InfoGain method.

**Table 3.** Classification accuracy rates (%) according to selected number of features for InfoGain method.

|  |  |  |
| --- | --- | --- |
| **Classification Methods** | **5 Features** | **10 Features** |
| Naive Bayes | 75.13 | 78.89 |
| Logistic Regression | 77.95 | 79.43 |
| SVM | 75.26 | 79.83 |
| KNN | 70.16 | 73.52 |
| Decision Tree | 79.03 | 80.10 |

**Table 4.** Classification accuracy rates (%) according to selected number of features for CFS method.

|  |  |  |
| --- | --- | --- |
| **Classification Methods** | **5 Features** | **10 Features** |
| Naive Bayes | 80.51 | 80.51 |
| Logistic Regression | 80.64 | 80.64 |
| SVM | 80.37 | 80.37 |
| KNN | 74.32 | 74.32 |
| Decision Tree | 79.03 | 79.03 |

**Table 5.** Classification accuracy rates (%) according to feature selection methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Methods** | **InfoGain** | **CFS** | **Wrapper** | **No Selection** |
| Naive Bayes | 78.89 | 80.51 | **81.04** | 76.47 |
| Logistic Regression | 79.43 | 80.64 | **83.87** | 83.06 |
| SVM | 79.83 | 80.37 | 82.93 | **84.40** |
| KNN | 73.52 | **74.32** | 73.79 | 72.17 |
| Decision Tree | **80.10** | 79.03 | 79.43 | 79.43 |

The results show that different image processing steps and feature extraction methods have an impact on classification accuracy rates. Best accuracy rate of 84.40% was achieved by SVM using all features. This result was achieved by using resizing and cropping of images, applying CLAHE, and edge-based segmentation.

In general, variations in accuracy rates have been observed as a result of combinations using different feature selection methods and classification methods. In some cases, a particular feature selection method provided better results for a particular classification method. Therefore, choosing the right feature selection method and classification method is important in terms of increasing the accuracy rates.

1. **Conclusion**

In this study, medical images of atelectasis and pneumonia diseases selected from the “ChestX-ray14” dataset were studied with various image processing and machine learning algorithms. The results show that SVM and Logistic Regression algorithms stand out in terms of accuracy in binary class classifications. Additionally, this study has shown that different feature selection methods and various numbers of selected feature features significantly change the accuracy result. This situation shows the effectiveness of classification methods on the data set.

This research enhances the current understanding of detecting and classifying atelectasis and pneumonia using X-ray images. The results obtained in this study may have significant implications for physicians to accurately diagnose these diseases. In addition, changing the image processing methods, determining a more suitable dataset, and increasing the selection of different numbers of features can increase the robustness of the model.

**References**

[1] Peroni, D., & Boner, A. (2000b). Atelectasis: mechanisms, diagnosis and management. *Paediatric Respiratory Reviews*, *1*(3), 274–278.

[2] Mackenzie, G. (2016b). The definition and classification of pneumonia. *Pneumonia*, *8*(1).

[3] K G, S., P, S., S, S. V., V, S., S, J., Mouna M, S., & T, S. (2022). Atelectasis, Cardiomegaly and Effusion Disease Prediction using Image Processing (Convolutional Neural Networks). 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA).

[4] Yee, S. L. K., & Raymond, W. J. K. (2020b). Pneumonia Diagnosis Using Chest X-ray Images and Machine Learning. *Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology*.

[5] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Papers with Code - ChestX-ray14 Dataset. (n.d.). Retrieved from https://paperswithcode.com/dataset/chestx-ray14

[7] The MathWorks Inc. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

[8] Frank, E., Hall, M. A., & Witten, I. H. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition.

1. \* Corresponding author. *e-mail address: hakan.aksebzeci@agu.edu.tr* [↑](#footnote-ref-2)