Babeş-Bolyai University
Faculty of Mathematics and Computer Science

Early Lung Cancer Detection Using Artificial Intelligence

Moghioroş Eric
Croitoru Andreea Bianca

Prof. Drd. Dioşan Laura
Drd. Limboi Sergiu

May 2025

Contents

1 Introduction **2**

2 Dataset **3**

2.1 Data Source 3

2.2 Data Description: Properties, Types and Formats 4

2.3 Data Annotation: Methods and Categories 4

2.4 Data Preprocessing: Augumentation and Cleaning 5

2.5 Data Splitting: Training, Test and Validation 6

2.6 Challenges and Limitations 7

1. Introduction

Lung cancer remains one of the deadliest cancers worldwide, with high mortality largely due to late-stage diagnosis. Timely and accurate detection is therefore critical for improving patient outcomes. Advances in artificial intelligence, particularly deep learning, have enabled the development of automated systems capable of analyzing medical images with performance approaching that of radiologists.

This project presents a multi-stage deep learning pipeline based on 2D Convolutional Neural Networks (CNNs) for the comprehensive analysis of chest CT scans. The goal is to assist in the detection, classification, and localization of lung cancer from individual axial slices of CT data.

The proposed system is composed of the following sequential models:

1. **Cancer Detection Model**: A binary classification CNN trained to determine whether a given CT slice contains cancerous tissue. This serves as the entry point of the diagnostic pipeline.
2. **Cancer Type Classification Model**: For slices identified as cancer-positive, a second CNN performs multi-class classification to predict the type of cancer, such as Non-Small Cell Lung Cancer (NSCLC), Small Cell Lung Cancer (SCLC), or other relevant subtypes.
3. **Tumor Localization Model**: A regression-based CNN designed to output bounding box coordinates that localize the suspected tumor region within the slice, providing spatial interpretability and aiding visual diagnosis.

Each model in the pipeline is trained and validated independently to ensure optimal performance for its specific task. By combining classification and localization, the system provides both diagnostic decisions and visual evidence to support clinical workflows. This modular approach also facilitates future improvements and scalability across different datasets and imaging protocols.

Future sections of this document will detail the dataset preprocessing steps, model architectures, training methodologies, evaluation metrics, and performance results.

2. Dataset

2.1 Data Source

The dataset used in this project is the *IQ-OTH/NCCD - Lung Cancer Dataset* (The Iraq-Oncol ogy Teaching Hospital/National Center for Cancer Diseases), publicly available from The Cancer Imaging Archive (TCIA). [1]

This dataset was collected in the above-mentioned specialist hospitals over a period of three months in fall 2019. It includes CT scans (originally collected in DICOM format) of patients diagnosed with lung cancer in different stages, as well as healthy subjects.

Each scan contains several slices (from 80 to 200), each of them represents an image of the human chest with different sides and angles. The cases vary in gender, age, educational attainment, area of residence, and living status. Some of them are employees of the Iraqi ministries of Transport and Oil, others are farmers and gainers. Most of them come from places in the middle region of Iraq, particularly, the provinces of Baghdad, Wasit, Diyala, Salahuddin, and Babylon.

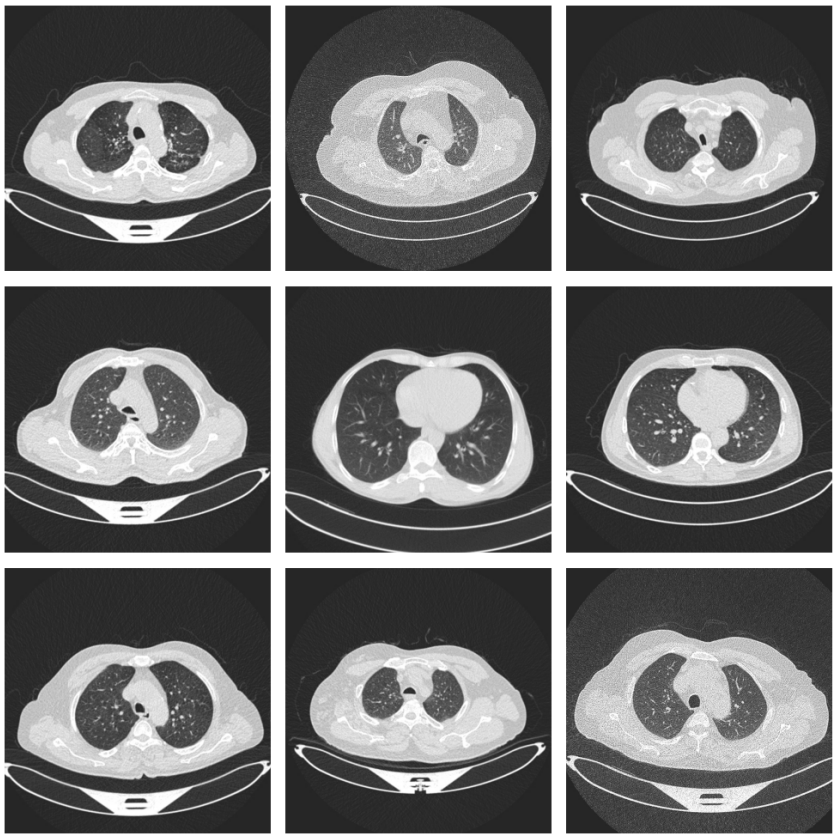

Figure 1. CT scan samples from dataset.

2.2 Data Description: Properties, Types and Formats

The IQ-OTH/NCCD lung cancer dataset consists of 3.609 chest CT scan images collected from 110 individual patients. Each image is stored in JPEG format with non-uniform resolutions.

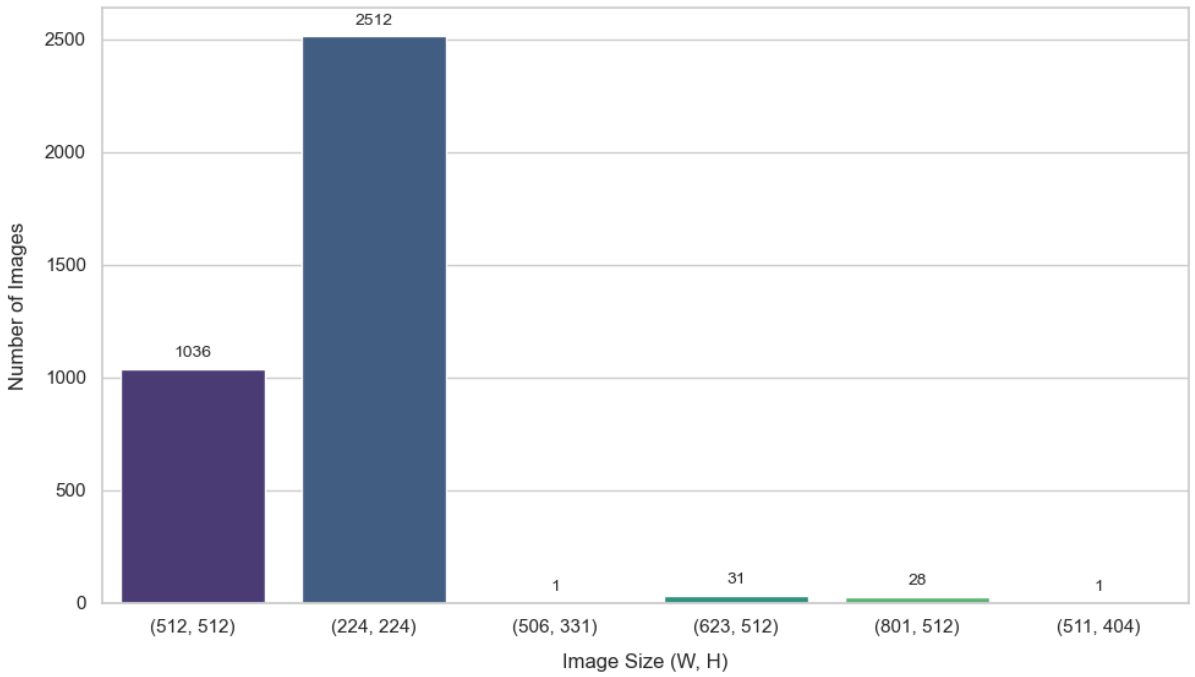

Figure 2. Distribution of images size.

2.3 Data Annotation: Methods and Categories

The dataset was manually organized into three distinct folders, each corresponding to a diagnostic category used for classification:

- **Normal** – images without visible nodules or abnormalities
- **Benign** – images containing nodules classified as non-cancerous
- **Malignant** – images containing nodules classified as cancerous

The labeling of each image was performed by a team of experienced oncologists and radiologists, ensuring a high degree of clinical accuracy and reliability in the annotation process. This expert-labeled dataset forms a strong foundation for training a supervised learning model in a sensitive medical context.

This folder-based organization enabled the model to correctly associate each image with its corresponding class label during training and evaluation. The distribution of these labeled samples is presented in Figure 3.

2.4 Data Preprocessing: Augumentation and Cleaning

As part of the preprocessing pipeline, non-square images were removed from the dataset prior to training. This decision was made to avoid potential issues associated with direct resizing or aspect-ratio-preserving padding, both of which can negatively impact the performance of CNN. To ensure consistency and compatibility with the input requirements of CNN, all remaining images were resized to a uniform resolution of 224×224 pixels.

Although techniques such as padding can preserve the original aspect ratio, they introduce artificial borders and non-informative regions into the image. These artifacts may be misinterpreted by the CNN as meaningful features, especially in the early convolutional layers, which are sensitive to spatial patterns. As a result, the model might learn irrelevant cues and generalize poorly, particularly in medical imaging tasks where precision is critical.

Furthermore, direct resizing of non-square images to a fixed square input size results in geometric distortion, altering the shape, size, and orientation of key anatomical structures such as lung nodules. This deformation can obscure clinically relevant patterns and lead to degraded model performance, especially in tasks involving subtle morphological differences between classes.

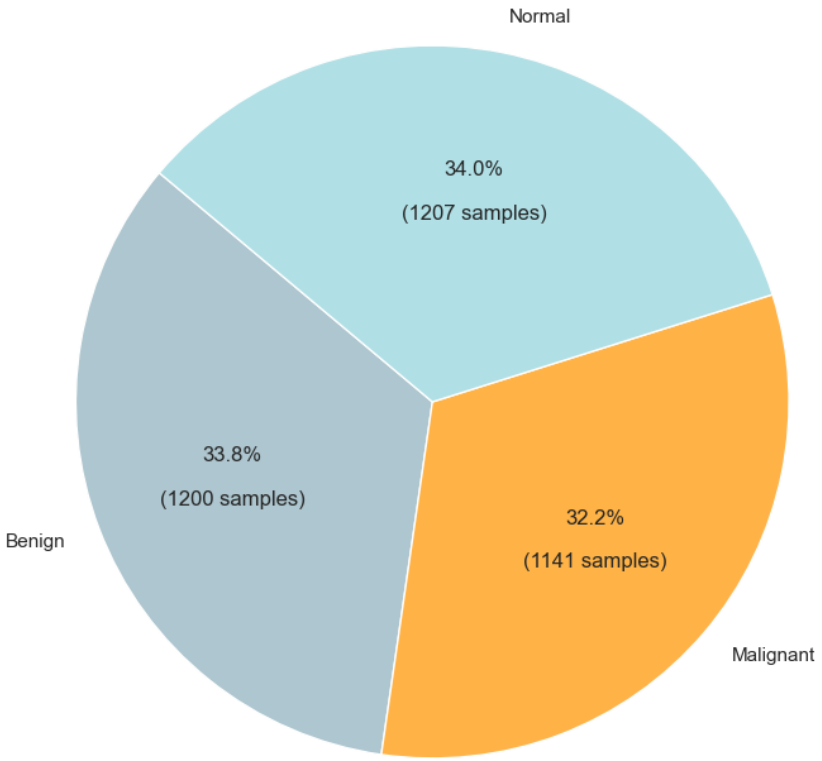

Figure 3. Distribution of images label after preprocessing.

The original dataset exhibited a significant imbalance among the diagnostic categories, which could adversely affect model training and lead to biased predictions. Specifically, the initial dataset included: 40 malignant cases, 15 benign cases and 55 normal (non-nodule) cases. To address this issue and ensure a more balanced distribution of classes, extensive data augmentation techniques to artificially expand the dataset were applied. The goal was to create a more uniform class distribution and enhance the model's ability to generalize across all categories. The augmentation techniques applied include: Horizontal Flip, Vertical Flip, Rotation, Colorjitter, Contour Crop, Gaussian Blur, Sharpeness, Contrast and Histogram Equalization [1].

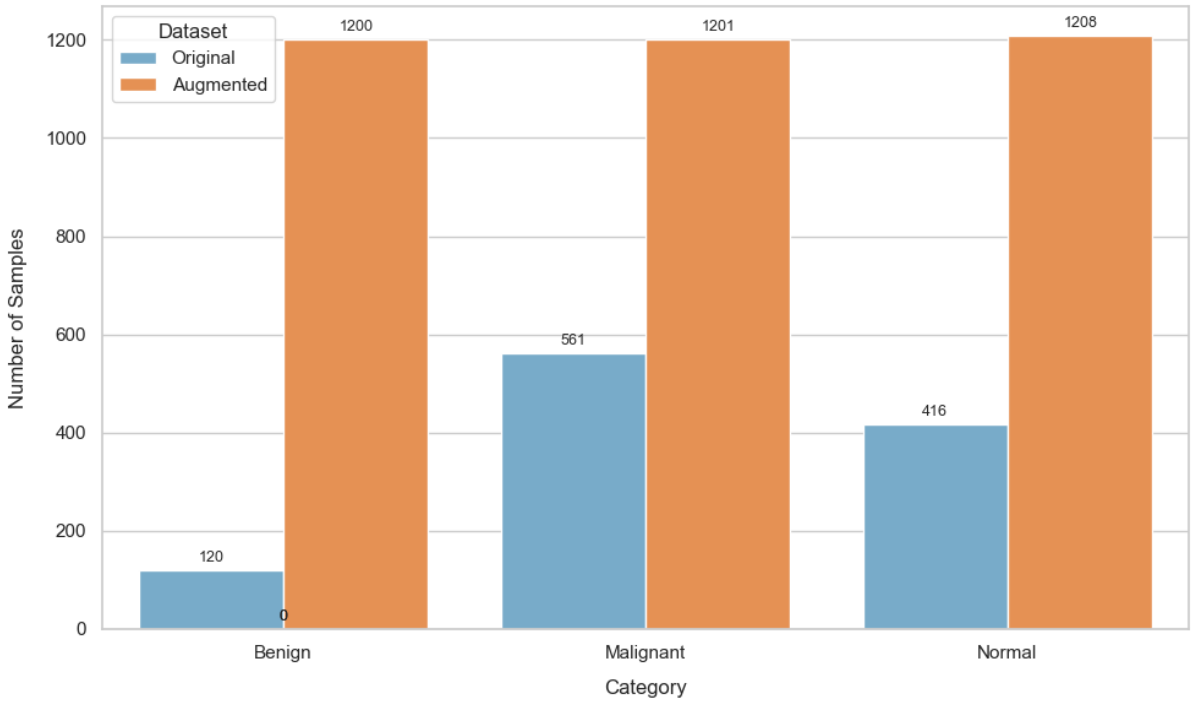

Figure 3. Distribution of images label compared to the original dataset.

2.5 Data Splitting: Training, Test and Validation

To ensure effective training, evaluation, and generalization of the Convolutional Neural Network (CNN), the dataset was split into three subsets using the following proportions: 75% Training Set, 15% Validation Set and 15% Test Set.

- **Effective Learning:** Allocating the majority of the data to the training set ensures that the model has sufficient samples to learn patterns, especially in a medical imaging context where inter-class variations may be subtle and complex.
- **Hyperparameter Tuning:** The validation set is used to monitor the model's performance during training and to fine-tune hyperparameters. This helps prevent

overfitting by evaluating how well the model generalizes to unseen data during the training process.

- **Unbiased Performance Evaluation:** The test set is strictly separated from the training and validation processes. It provides an objective estimate of the model's real-world performance, helping ensure that performance metrics are not inflated by exposure to the training data.

This division plays a crucial role in the training workflow and contributes to the development of a robust and unbiased model.

2.6 Challenges and Limitations

References

[1] Subhajeet Das. *IQ-OTH/NCCD Lung Cancer Dataset (Augmented)*. 2025. DOI: 10.34740/KAGGLE/DS/6582139. URL: https://www.kaggle.com/ds/6582139.