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

Lung cancer is one of the most lethal cancers globally, with a high mortality rate primarily caused by late detection. Early and accurate diagnosis plays a pivotal role in improving treatment outcomes and patient survival. Recent advances in artificial intelligence, particularly in deep learning, have paved the way for automated systems that can assist in medical image analysis with a performance nearing that of expert radiologists.

This project introduces a multi-stage deep learning pipeline designed to support the detection, classification, and localization of lung abnormalities in axial slices from chest CT scans. The pipeline employs 2D Convolutional Neural Networks (CNNs) in a modular architecture to replicate the diagnostic process used by clinicians.

The pipeline is composed of the following key components:

1. **Primary Classification Model:** A multi-class CNN that analyzes a given CT slice and classifies it into one of three categories: *Normal*, *Benign*, or *Malignant*. This model acts as the entry point of the pipeline and quickly filters out normal cases.
2. **Tumor Localization Model:** If the classification result is *Benign* or *Malignant*, the image is passed to a secondary CNN that performs object localization. This model predicts bounding box coordinates that delineate the tumor area within the image, thus providing spatial context and visual support for clinical interpretation.

By combining classification and localization, the system not only flags potentially pathological cases but also highlights the specific region of interest, making it a valuable decision-support tool for radiologists. Each model is trained and validated independently to optimize its performance and ensure robustness across diverse imaging conditions.

This modular design allows for flexible updates, such as replacing or fine-tuning individual components as more data becomes available or clinical needs evolve. Future sections will describe the dataset, preprocessing techniques, augmentation strategies, model architectures, training methodologies, evaluation metrics, and experimental results.

2. Dataset

2.1 Data Source

The dataset used in this project is the IQ-OTH/NCCD Lung Cancer Dataset (The Iraq-Oncology 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 representing an image of the human chest captured from different sides and angles. The patient cases exhibit diversity in gender, age, educational background, area of residence, and occupation — including ministry employees, farmers, and laborers — mostly from the central region of Iraq (Baghdad, Wasit, Diyala, Salahuddin, Babylon).

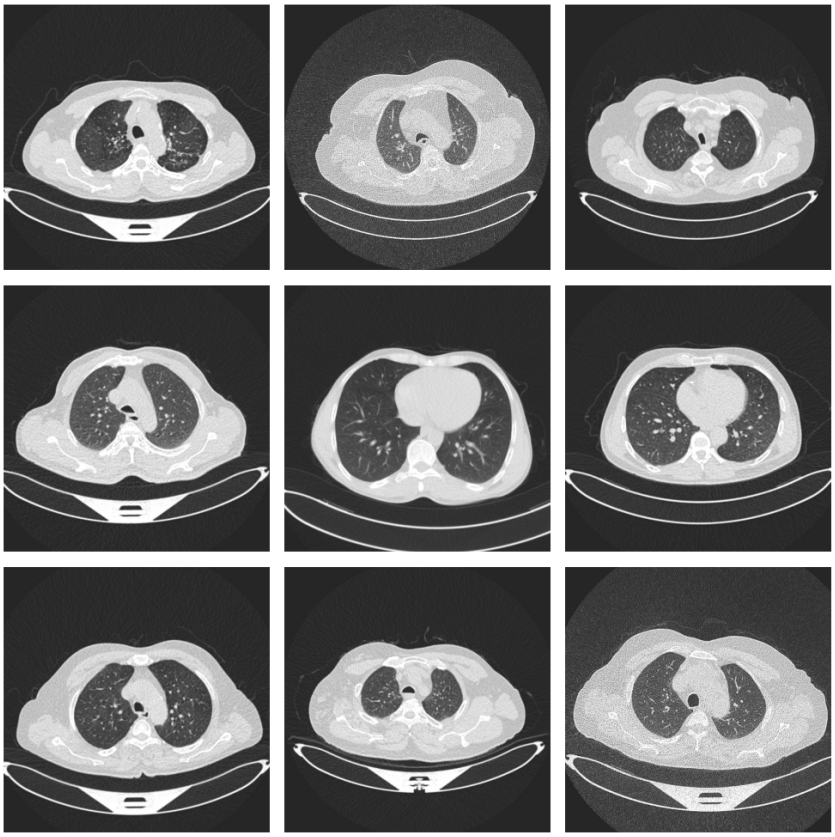

Figure 1. CT scan samples from dataset.

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 and exhibits non-uniform resolutions.

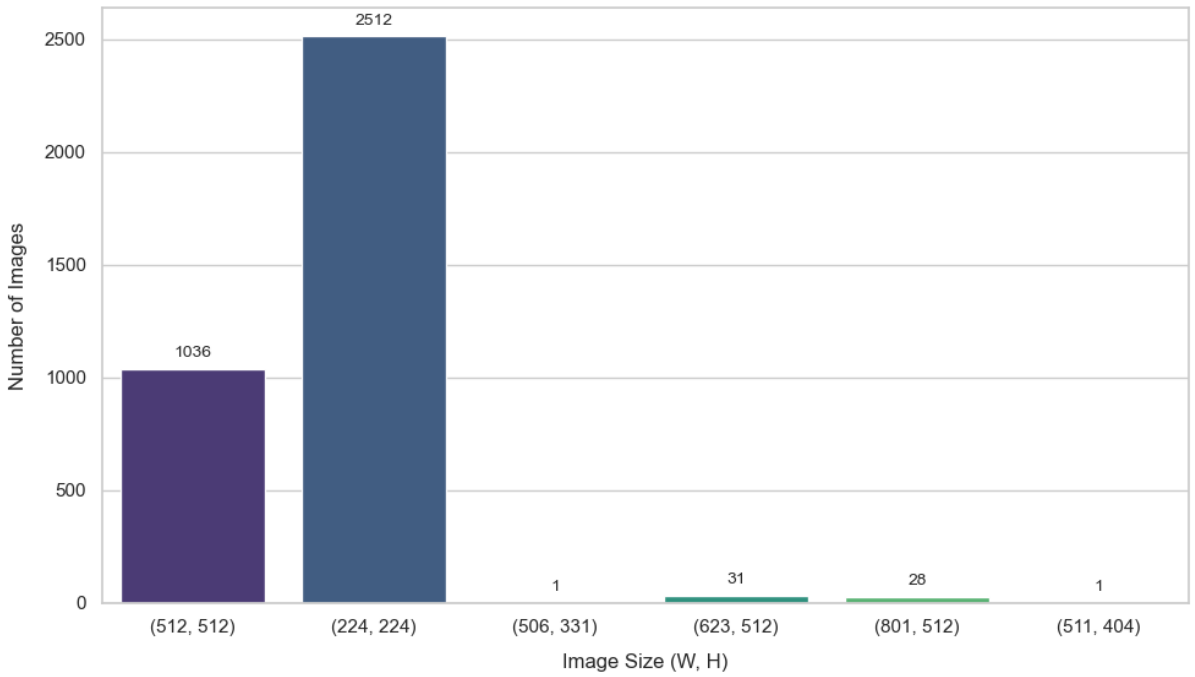

Figure 2. Distribution of image sizes.

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.

Labeling was conducted by a team of expert oncologists and radiologists, ensuring a high level of clinical accuracy. This expert-labeled structure provides a reliable foundation for training a supervised learning model in a sensitive medical context.

The folder-based annotation format directly facilitates class-based model training. Class distribution after preprocessing is shown in Figure 3.

2.4 Data Preprocessing: Augmentation and Cleaning

To standardize the dataset for deep learning input, all non-square images were removed to avoid issues arising from distortion due to resizing or padding. Remaining images were resized to a uniform 224×224 resolution, compatible with CNN input requirements.

Techniques like padding — though preserving aspect ratio — introduce artificial borders, which CNNs may misinterpret as features. Likewise, direct resizing of non-square images introduces distortions that can obscure anatomical details crucial for accurate medical interpretation.

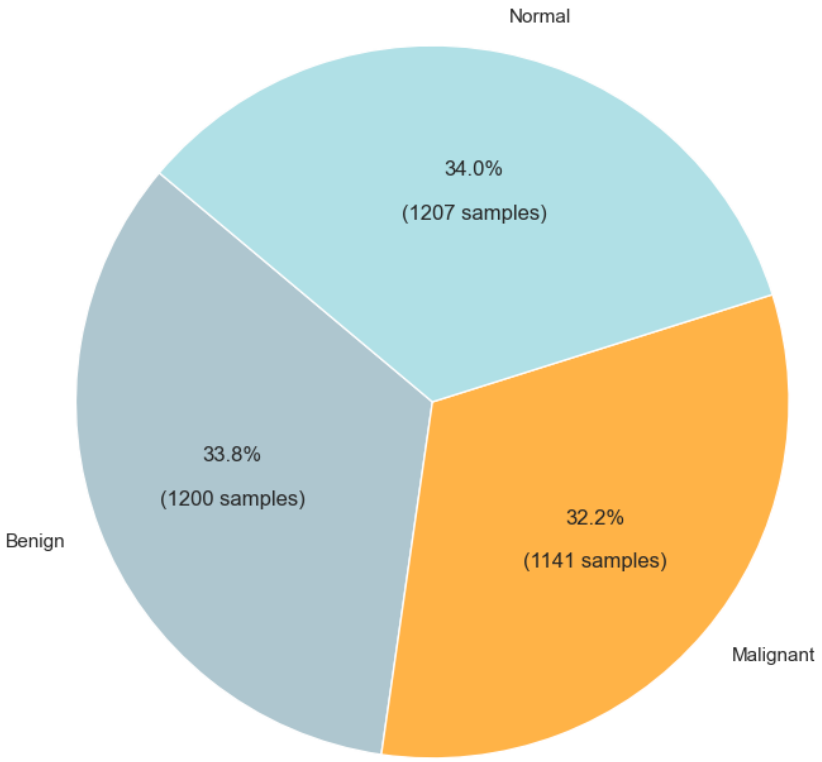

Figure 3. Distribution of image labels after preprocessing.

Given the dataset's initial class imbalance (40 malignant, 15 benign, and 55 normal), extensive augmentation techniques were used to generate additional examples, particularly for underrepresented classes. The augmentation strategies included: Horizontal Flip, Vertical Flip, Rotation, Colorjitter, Contour Crop, Gaussian Blur, Sharpness Adjustment, Contrast Manipulation, and Histogram Equalization [1].

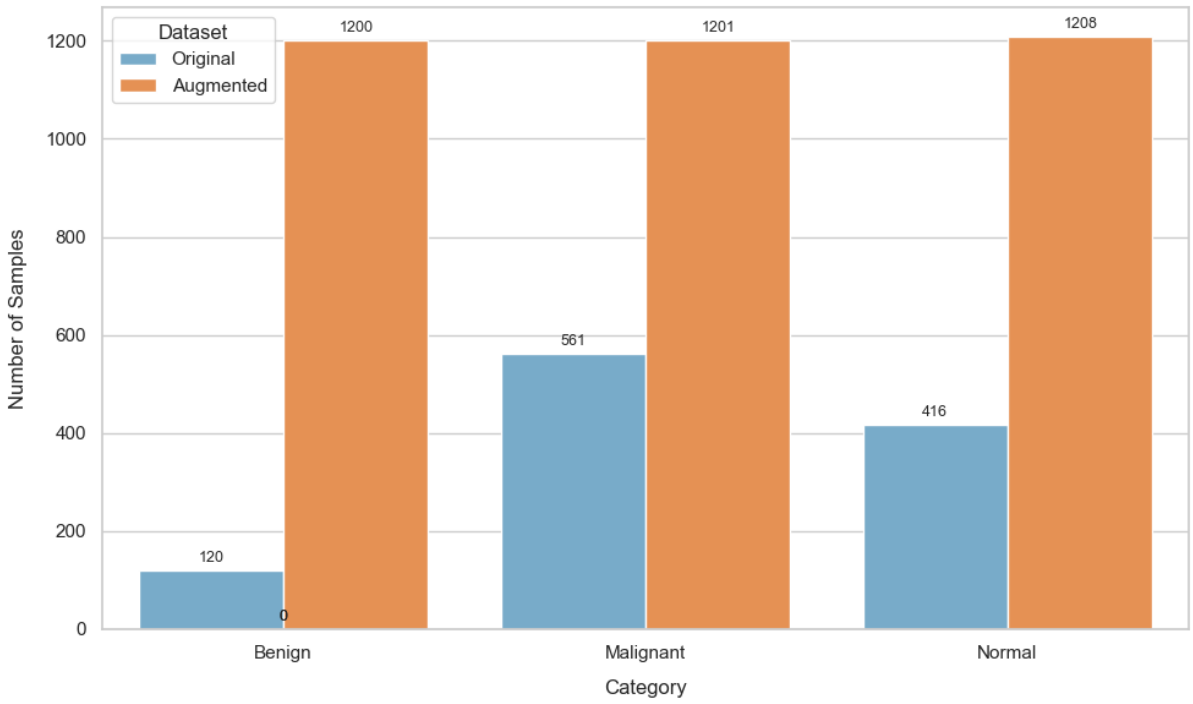

Figure 4. Distribution of image labels compared to original dataset.

2.5 Data Splitting: Training, Test and Validation

To develop and evaluate the CNN effectively, the dataset was divided into three subsets with the following proportions: 75% for training, 15% for validation, and 15% for testing.

- **Effective Learning:** The large training portion ensures enough samples for learning complex patterns, especially in high-variance medical data.
- **Hyperparameter Tuning:** The validation set supports model tuning and early stopping by evaluating generalization during training.
- **Unbiased Evaluation:** The test set remains untouched during training, providing an objective measure of real-world performance.

This structured split is essential for developing a generalizable and clinically reliable model.

2.6 Challenges and Limitations

Despite its usefulness, the IQ-OTH/NCCD dataset presents several limitations:

- **Limited Dataset Size:** Only 3,609 CT images from 110 patients — a relatively small dataset — may lead to overfitting and reduced generalization.
- **Class Imbalance:** A severe initial imbalance, especially in benign cases, may affect

model reliability, despite augmentation efforts.

- **Annotation Variability:** Expert annotations can still be subjective, with inter-observer variability potentially introducing label noise.
- **Lack of 3D Context:** Working with individual 2D slices disregards spatial continuity across scans, possibly omitting valuable diagnostic information.
- **Resolution Inconsistency:** JPEG compression artifacts and varying image quality hinder feature consistency and model accuracy.
- **Preprocessing Trade-offs:** Removal of non-square images and resizing may discard informative content or distort lesion morphology.
- **Geographic and Demographic Bias:** Patients from a specific region in Iraq limit generalizability across different populations and scanner environments.

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.