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

3 Model Architecture **8**

3.1 Overview of CNN structure (layers, activations) 8

3.2 Input/output format 10

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-Oncol ogy Teaching Hospital/National Center for Cancer Diseases), publicly available from The Cancer Imaging Archive (TCIA). [5]

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.

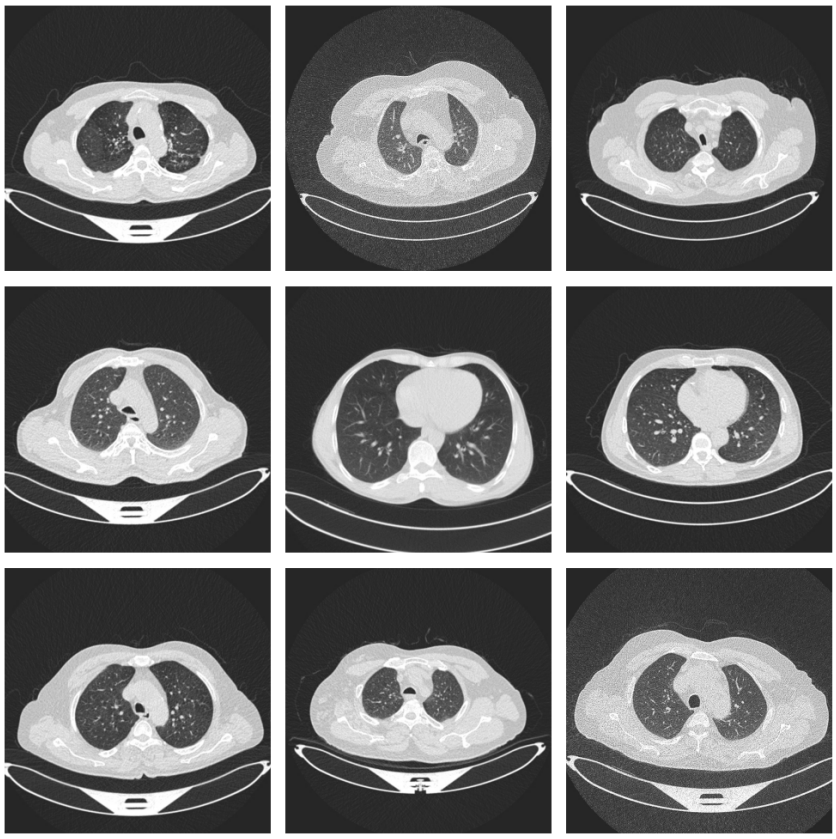


Figure 1. CT scan samples from dataset.

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.

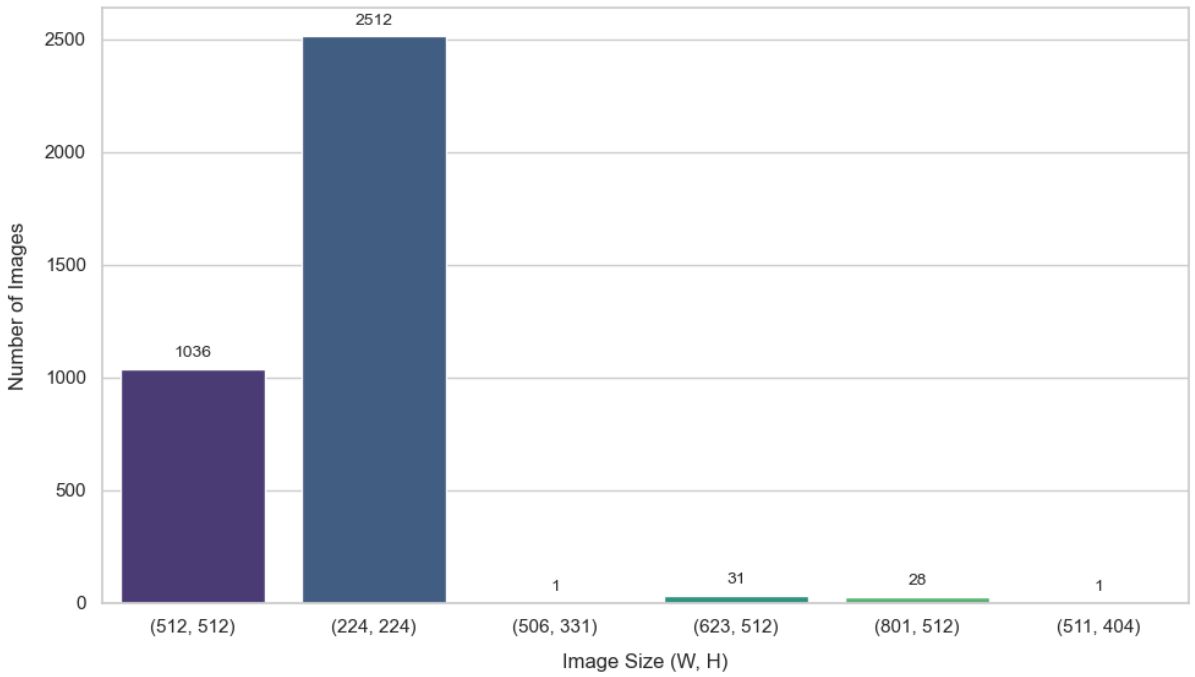


Figure 2. Distribution of images size.

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.

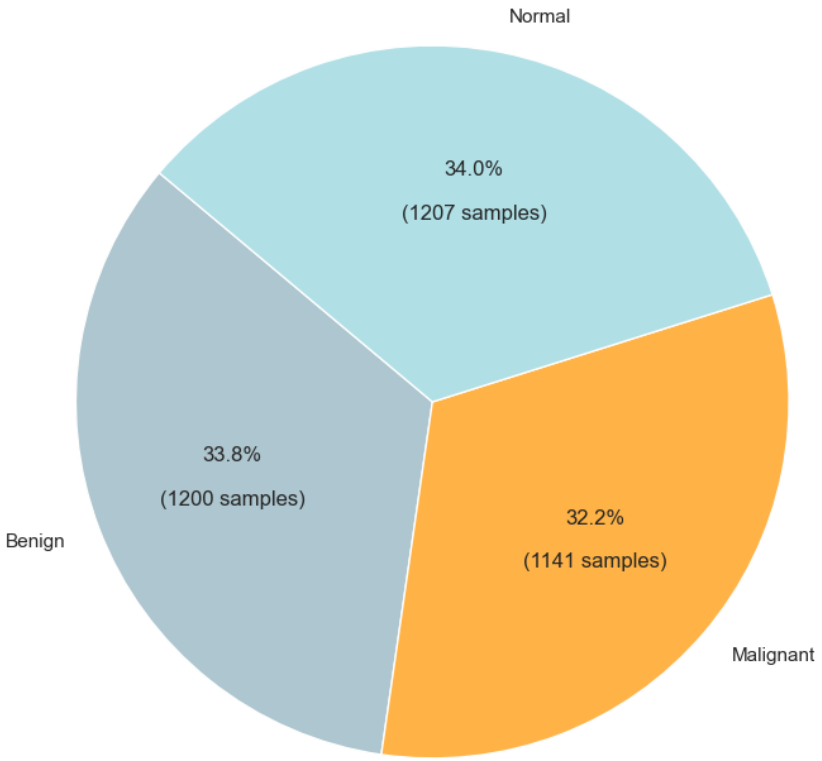


Figure 3. Distribution of images label after preprocessing.

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 [5].

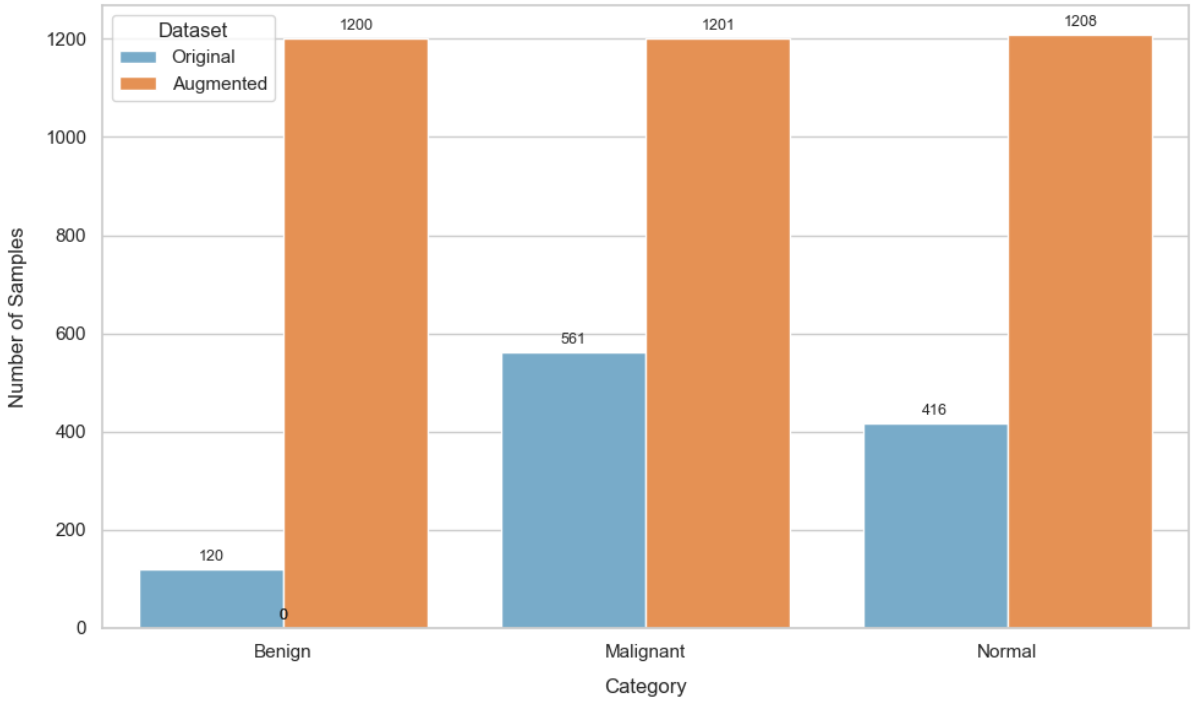


Figure 3. Distribution of images label compared to the original dataset.

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

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.

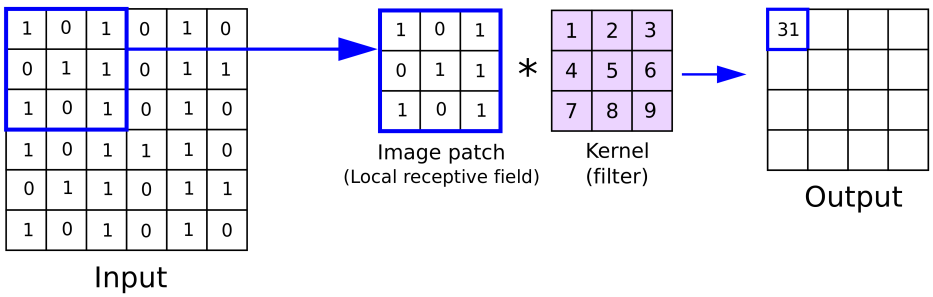
3. Model Architecture

Lung cancer detection using **Convolutional Neural Networks (CNNs)** has become a promising approach in medical imaging due to CNNs' ability to automatically learn and extract relevant features from complex CT scan images.

3.1 Overview of CNN structure (layers, activations)

- **Convolutional Layers:** These layers form the core of the CNN architecture. Each convolutional layer applies multiple filters (kernels) to the input or previous layer's feature maps to detect local patterns such as edges, textures, and shapes relevant to lung cancer.
  - **Filter size:** Typically 3×3 filters are used because they strike a balance between capturing fine-grained details and computational efficiency. [3]
  - **Number of filters:** The number of filters usually increases progressively through the network (e.g., 32, 64, 128, 256), allowing the model to learn from simple to complex features hierarchically.
  - **Stacked convolutions:** Multiple convolutional layers are stacked to deepen the network, enabling the detection of subtle malignancies and complex lung tissue patterns.

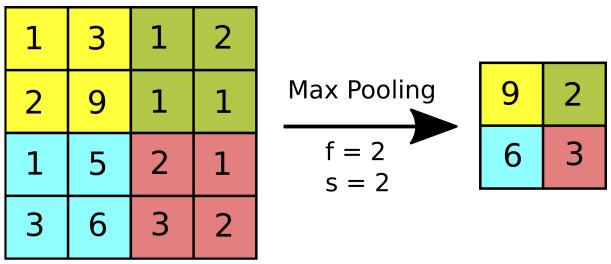
  The convolutional layers are vital for extracting meaningful features from the CT scans, which are critical for distinguishing between normal, benign, and malignant tissues.



Convolutional layer representation. [1]

- **Batch Normalization:** After each convolutional operation, batch normalization layers are inserted to normalize the activations. This technique stabilizes and accelerates training by reducing internal covariate shift, which helps the model converge faster and generalize better on unseen data. [3]

- **Activation Functions:** The Rectified Linear Unit (ReLU) activation function is applied after batch normalization in each convolutional block. ReLU introduces non-linearity by outputting zero for negative inputs and the input itself for positive values.
  - **Impact:** ReLU prevents the vanishing gradient problem, allowing deeper networks to learn effectively.
  - **Training efficiency:** It speeds up convergence during training.
  - **Feature learning:** Enables the network to learn complex, non-linear features essential for accurate lung cancer classification. [3]
- **Pooling Layers:** Max pooling layers with a typical pool size of 2×2 are inserted after convolutional blocks to downsample the feature maps.
  - **Purpose:** Reduce spatial dimensions, lowering computational load.
  - **Robustness:** Provides translation invariance, making the model less sensitive to small shifts or distortions in lung images.
  - **Information retention:** Max pooling preserves the most prominent features in each region, which is crucial for detecting cancerous nodules.



Max pooling representation. [1]

- **Dropout Layers:** Dropout layers randomly deactivate a fraction of neurons during training (commonly 0.5 dropout rate). This prevents the network from overfitting by forcing it to learn redundant and more robust feature representations. These layers are particularly important in medical imaging, where datasets may be limited and overfitting is a risk.
- **Flatten Layer:** After the convolutional and pooling layers, the multi-dimensional feature maps are flattened into a one-dimensional vector. This transformation prepares the data for the fully connected layers that perform the final classification.
- **Dense Layer:** One or more fully connected layers follow the flattening step. These layers integrate all extracted features to make a final decision on the presence and

type of lung cancer. Dense layers typically contain hundreds of neurons to capture complex feature interactions.

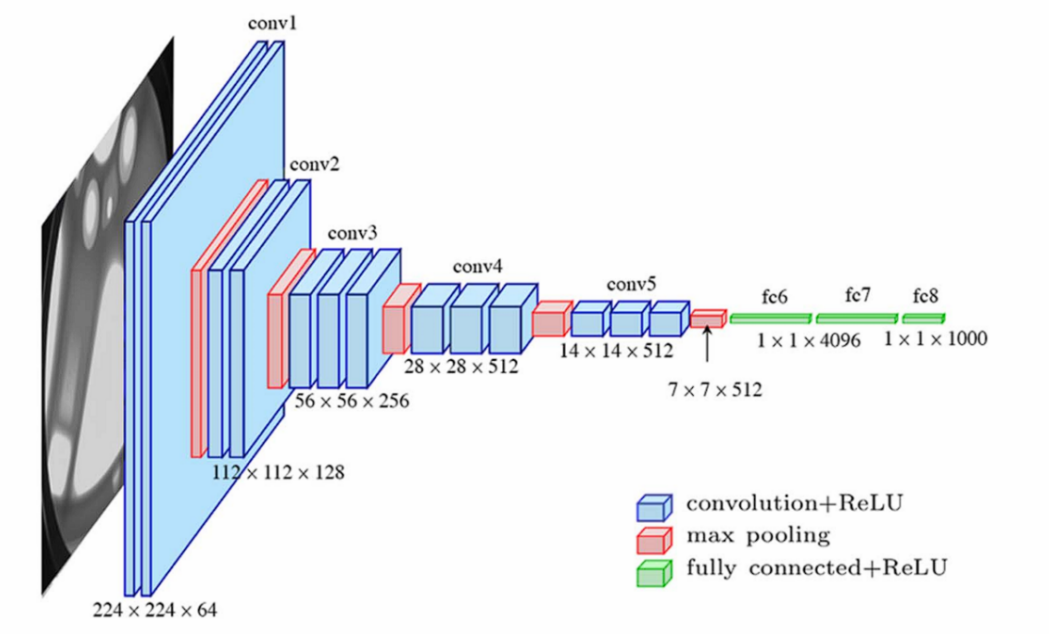


Diagram of Very Deep Convolutional Networks. [2]

3.2 Input/output format

- **Input Layer and Preprocessing:** The CNN model begins with an input layer designed to accept lung CT scan images resized to a standardized dimension, often 256×256 pixels. This resizing ensures consistent input size for the network and preserves critical spatial information necessary for detecting lung nodules or tumors. Preprocessing steps prior to input include noise reduction (techniques such as thresholding and segmentation are applied to remove irrelevant regions like bones, air, and background noise, isolating lung tissues for focused analysisand) and contrast enhancement to improve image quality while avoiding loss of important features. [4]
- **Output Layer and Prediction:**
  - **Binary classification:**
    - **Structure:** A single neuron with a sigmoid activation (e.g., cancerous vs. non-cancerous).
    - **Output:** A scalar value between 0 and 1 representing the probability that the input image contains cancerous tissue. A threshold (commonly

0.5) is applied to convert the probability into a class label (e.g., benign if probability less than 0.5, malignant otherwise).

- **Multiclass classification:**
  - **Structure:** Multiple neurons equal to the number of classes (e.g., normal, benign, malignant) with a softmax activation function.
  - **Output:** A probability distribution over all classes, where the sum of probabilities equals 1. The class with the highest probability is selected as the predicted label. For example, if the output is [0.1, 0.7, 0.2], the model predicts "benign" with 70% confidence.
- **Confidence Scores:** The output probabilities provide a measure of confidence in the prediction. This is essential for clinical applications where uncertain cases may require further examination or additional testing. Well-calibrated models produce probabilities that accurately reflect true likelihoods. Calibration techniques such as Platt scaling or isotonic regression may be applied post-training to improve reliability.

References

[1] Anh. Reynolds. *Convolutional Neural Networks (CNNs)*. 2019. URL: https://anhreynolds.com/blogs/cnn.html.

[2] Gaudenz Boesch. *Very Deep Convolutional Networks (VGG) Essential Guide*. 2021. URL: https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/.

[3] Hammad, M., ElAffendi, M., El-Latif, A.A.A. et al. "Explainable AI for lung cancer detection via a custom CNN on CT images". In: *Sci Rep* 15.12707 (2025).

[4] Pathan, Sameena et al. "An optimized convolutional neural network architecture for lung cancer detection." In: *APL bioengineering* 8.2 (2024). DOI: 10.1063/5.0208520.

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