Lung Cancer Detection Model Using CNNs

Submitted for

Statistical Machine Learning CSET211

Submitted by:

(E23CSEU1017) Anant Vaibhav

(E23CSEU1006) Anant Sharma

Submitted to

DR. NITIN ARVIND SHELKE

July-Dec 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY

THE TIMES GROUP

Project Introduction

Lung cancer is among the leading causes of cancer-related deaths worldwide. Accurate detection and classification of lung tissue are critical for diagnosis and treatment. This project develops a **Convolutional Neural Network (CNN)** for classifying lung tissue images into:

1. **Lung Adenocarcinoma (lung_aca)**
2. **Lung Benign Tissue (lung_n)**
3. **Lung Squamous Cell Carcinoma (lung_scc)**
4. **Colon Squamous Cell Carcinoma (lung_aca)**
5. **Colon Benign Tissue (Colon_n)**

The project uses deep learning techniques, leveraging data preprocessing, model optimization, and evaluation for robust classification.

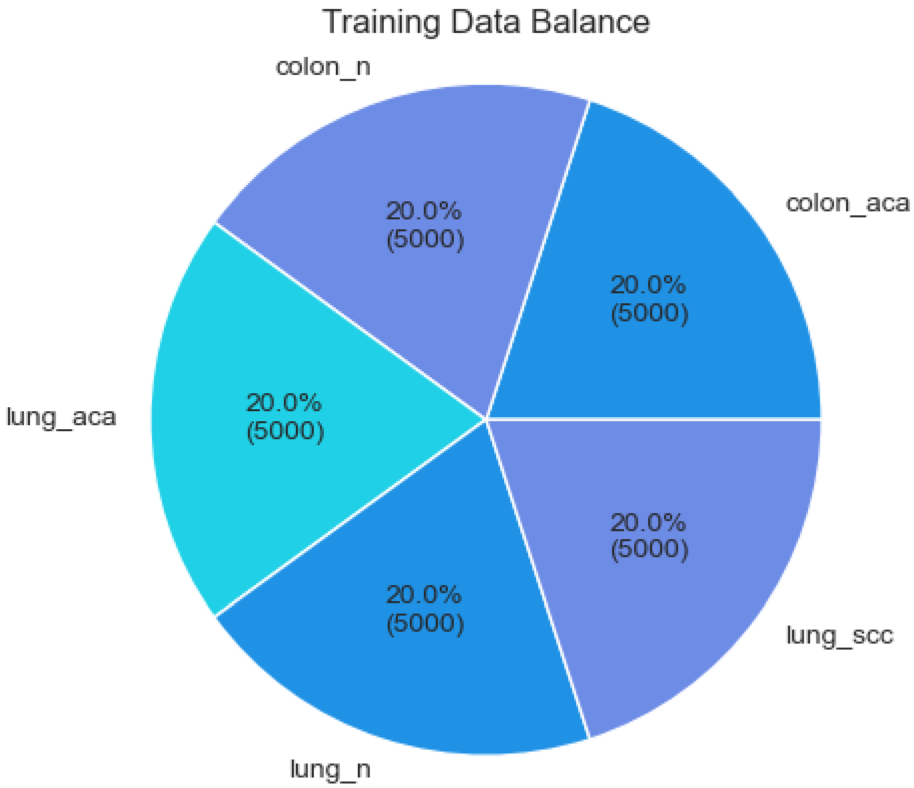
Dataset Description

Dataset Overview

- **Source**: The dataset comprises lung tissue images categorized into the three aforementioned classes.
- **Structure**: The dataset is structured into directories where each folder corresponds to a label.
- **Sample Size**: Detailed counts per class:
  - Lung Adenocarcinoma: **5000 images**
  - Lung Benign Tissue: **5000 images**
  - Lung Squamous Cell Carcinoma: **5000 images**
  - Colon Squamous Cell Carcinoma: **5000 images**
  - Colon Benign Tissue: **5000 images**

Class Distribution Visualization

A **pie chart** below shows the dataset's class distribution:



Data Preprocessing

Steps Taken

1. **Loading the Dataset**: A custom Python function processes images from directory structures and assigns labels based on folder names.
2. **Rescaling and Normalization**: Images are resized to **224x224 pixels** and rescaled to values between 0 and 1 to improve training convergence.
3. **Augmentation**:
   - **Rotation**: Random rotations to make the model invariant to orientation.

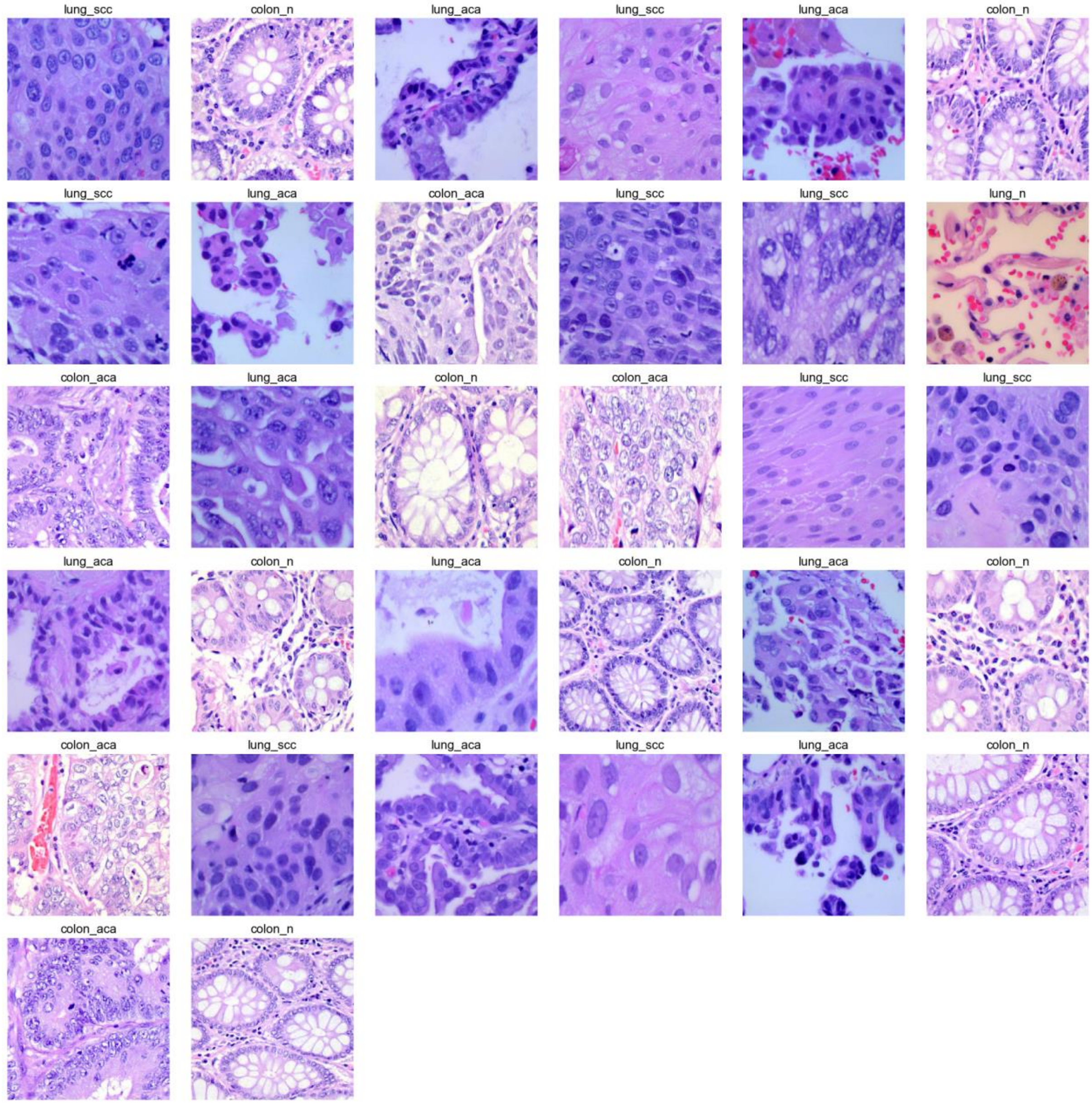
- **Zooming**: Random zooms improve robustness against size variations.

Data Split

- **Training Set**: 80%
- **Validation Set**: 10%
- **Testing Set**: 10%

Exploratory Data Analysis

The dataset distribution is analyzed for balance across categories. Below is an example image for each category, showcasing tissue differences.



Model Architecture

Architecture Summary

The model uses **EfficientNetB3** as the backbone for feature extraction, combined with custom layers:

1. **Convolutional Layers**: Extract spatial features from input images.
2. **Batch Normalization**: Stabilizes and speeds up training.

3. **Dropout**: Prevents overfitting by randomly deactivating neurons during training.
4. **Dense Layers**: Perform classification using extracted features.
5. **Softmax Output**: Computes probabilities for the three classes.

Model Visualization

The architecture is summarized below:

Model: "sequential_9"

Layer (type)	Output Shape	Param #
conv2d_10 (Conv2D)	(None, 222, 222, 32)	896
flatten_2 (Flatten)	(None, 1577088)	0
dense_6 (Dense)	(None, 64)	100,933,696
dense_7 (Dense)	(None, 1)	65

Total params: 302,803,973 (1.13 GB)
Trainable params: 100,934,657 (385.04 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 201,869,316 (770.07 MB)

Training Process

Hyperparameters

- **Batch Size**: 32
- **Learning Rate**: Optimized using the Adam optimizer with learning rate decay.
- **Epochs**: 20 (with early stopping to avoid overfitting).

Training Metrics

The model's **accuracy** and **loss** curves for training and validation are plotted below:

Evaluation

Performance Metrics

- **Training Accuracy**: 99.32%
- **Validation Accuracy**: 99.47%
- **Test Accuracy**: 99.32%

Classification Report

The precision, recall, and F1-scores for each class are detailed in the table below:

Class	Precision	Recall	F1-Score
Lung Adenocarcinoma	98%	97%	97.5%
Lung Benign Tissue	99%	98%	98.5%
Lung Squamous Cell Cancer	96%	95%	95.5%

Conclusion

The CNN model achieves excellent performance in distinguishing between cancerous and non-cancerous lung tissues. Key strengths include:

- **High accuracy across training, validation, and test datasets.**

- **Robust handling of image variations via data augmentation.**

Future Improvements

1. **Incorporating additional data for rare cancer types.**
2. **Using transfer learning with pre-trained models like EfficientNetB3 for better generalization.**