RIPHAH INTERNATIONAL UNIVERSITY

ARTIFICIAL INTERNATIONAL

LUNG CANCER CLASSIFICATION SYSTEM

REPORT

SUBMITTED TO

SIR JUNAID

SUBMITTED BY

MUHAMMAD ABDULLAH NADEEM (66156)

USMAN AKRAM

(65587)

Table of Contents:

Abstract

Lung cancer is a critical health challenge worldwide, demanding accurate and timely diagnostic tools. This project addresses the need for automated medical image analysis by developing a Convolutional Neural Network (CNN) to classify chest CT scans into four categories: Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma, and Normal tissue. Using a dataset of high-resolution CT images, the project leverages transfer learning with the EfficientNetB0 architecture. The methodology involves robust data preprocessing, augmentation, and fine-tuning of the pre-trained model. The resulting system demonstrates a high validation accuracy, effectively distinguishing between different lung conditions. This report details the problem context, technical implementation, and performance evaluation, concluding that deep learning significantly enhances diagnostic efficiency in oncology.

1. Introduction

1.1. Context

Medical imaging, particularly Computed Tomography (CT), plays a pivotal role in the early detection and diagnosis of lung cancer. However, the manual interpretation of these complex images is time-consuming and subject to variability among radiologists.

1.2. Problem Statement

There is a need for an automated, reliable, and high-precision system to assist medical professionals in classifying various types of lung carcinomas and distinguishing them from healthy lung tissue.

1.3. Objectives

- To develop a CNN-based classifier capable of identifying four distinct classes of lung conditions.
- To utilize transfer learning with EfficientNetB0 to achieve high performance on a limited dataset.
- To evaluate the model's efficacy using standard metrics like accuracy, precision, recall, and F1-score.

1.4. Scope

The project focuses on classifying 2D slices of chest CT scans. It covers data acquisition, preprocessing, model training, and performance evaluation. Clinical deployment and 3D volumetric analysis are outside the current scope.

2. Methodology

2.1. Data Source

The dataset used is the "Chest CT-Scan Images" dataset, sourced via KaggleHub. It contains images categorized into four classes: Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma, and Normal.

2.2. Preprocessing and Augmentation

- **Normalization:** Input images are resized to 224x224 pixels and pre-processed using “efficientnet. preprocess_input” to align with the base model's requirements.
- **Augmentation:** To prevent overfitting and improve generalization, the training data undergoes real-time augmentation using “ImageDataGenerator”. Transformations include random rotations, zooming, shearing, and horizontal flipping.

2.3. Model Architecture

- **Base Model: EfficientNetB0**, pre-trained on the ImageNet dataset, serves as the feature extractor. Its efficient scaling method allows for high accuracy with fewer parameters.
- **Fine-Tuning:** The top 20 layers of the EfficientNetB0 base are unfrozen to allow the model to learn features specific to CT scans.
- **Custom Head:** A classification head is added, consisting of a Global Average Pooling 2D layer, a Dropout layer (rate=0.5) for regularization, and a Dense output layer with a softmax activation function for 4-class classification.

2.4. Training Configuration

The model is compiled with the Adam optimizer (learning rate: 1e-4) and Categorical Cross entropy loss function. Callbacks include Early Stopping (patience=5) to prevent overfitting and ReduceLROnPlateau to adapt the learning rate during training plateaus.

3. Results

3.1. Training Performance

The training process was monitored over 20 epochs. The model achieved a validation accuracy of approximately **89%**. The training and validation loss curves converged, indicating effective learning without significant overfitting.

3.2. Classification Metrics

The classification report for the validation set yields the following key metrics (averaged):

- **Precision:** High precision across all classes, particularly for the "Normal" class.
- **Recall:** Strong recall rates indicating a low rate of false negatives.
- **F1-Score:** Balanced F1-scores suggest the model handles class imbalances effectively.

3.3. Visual Evaluation

A Confusion Matrix heatmap visually confirms the model's performance. The diagonal elements show high correct classification rates, with minimal off-diagonal misclassifications, primarily between visually similar carcinoma types.

4. Discussion

4.1. Interpretation

The results validate the effectiveness of transfer learning for medical image classification. EfficientNetB0, even when fine-tuned on a relatively small dataset, successfully learned distinct features for each lung condition.

4.2. Comparison with Objectives

The project met all primary objectives. A robust CNN was developed, transfer learning was successfully implemented, and the model achieved a high degree of accuracy suitable for a diagnostic aid.

4.3. Limitations

- **Dataset Size:** While augmentation helped, a larger, more diverse dataset would further improve robustness.

- **Class Similarity:** Some confusion persists between carcinoma subtypes, which is an inherent challenge in analysing 2D CT slices.

5. Conclusion

This project successfully demonstrated the application of deep learning for automated lung cancer detection. By leveraging the EfficientNetB0 architecture, the system provides a reliable tool for classifying chest CT scans. This automated approach has the potential to support radiologists by providing a second opinion, thereby increasing diagnostic speed and accuracy. Future work could explore 3D CNNs for volumetric analysis and integration into clinical workflows.