Lungs cancer classification using Convolutional Neural Network (CNN)

Tayyba Salamat Sana Akbar Ubaida Waheed

Introduction:

Lung cancer is the leading cause of cancer-related deaths worldwide, and early detection is crucial for treatment and survival. One promising approach for lung cancer classification is using convolutional neural networks (CNN), which is a type of deep learning model.

A CNN is a type of deep learning model that is particularly well-suited for image classification tasks because they are able to automatically learn features from images, making them well-suited to image classification tasks. It consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers are responsible for detecting features in the input image while pooling layers are used for downsampling the feature maps. Fully connected layers are used for classification.

Literature Review:

Several studies have reported the use of CNNs for lung cancer type classification. One study used a CNN model trained on a dataset of computed tomography (CT) scans to classify lung cancer into three subtypes: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The model achieved an accuracy of 96.6%, which is a significant improvement over traditional methods such as radiologists' diagnoses.One study published in the journal "Cancer Imaging" used a CNN to classify lung nodules as benign or malignant on CT scans with an accuracy of 96%. Another study in the "Journal of Medical Imaging" used a CNN to classify lung cancer on PET-CT images with an accuracy of 91%.Additionally, there are pre-trained models available like ChexNet, which is a 121-layer CNN trained on chest radiographs and has been shown to perform well in detecting lung pathology.

Methodology:

About Data Set:

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) lung cancer dataset was collected in the above-mentioned specialist hospitals over a period of three months in fall 2019. It includes CT scans of patients diagnosed with lung cancer in different stages, as well as healthy subjects. IQ-OTH/NCCD slides were marked by oncologists and radiologists in these two centers. The dataset contains a total of 1097 images .These cases are grouped into three classes: normal, benign, and malignant. The CT scans were originally collected in DICOM format(Digital Imaging and Communications in Medicine standard). All images were de-identified before performing analysis. Written consent was waived by the oversight review board. The study was approved by the institutional review board of participating medical centers. Each scan contains several slices. The number of these slices range from 80 to 200 slices, each of them represents an image of

the human chest with different sides and angles. The 110 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.

About Model:

A CNN model was trained using this data and the architecture of the model was optimized using techniques such as dropout, Adam optimizer and early stopping.

The model was then validated using a separate test dataset.The first preprocessing of the dataset is done by rescaling and resizing. Rescaling is done by making all the pixel values between 0 and 1 while resizing is done by changing image size to 256x256.Then we split it into the train, validate which is 80%, 20%, respectively, test data of 60 images(20 of each class) is separated at the start and made them shuffle randomly. We made batches of size 81. And images size is 256x256.

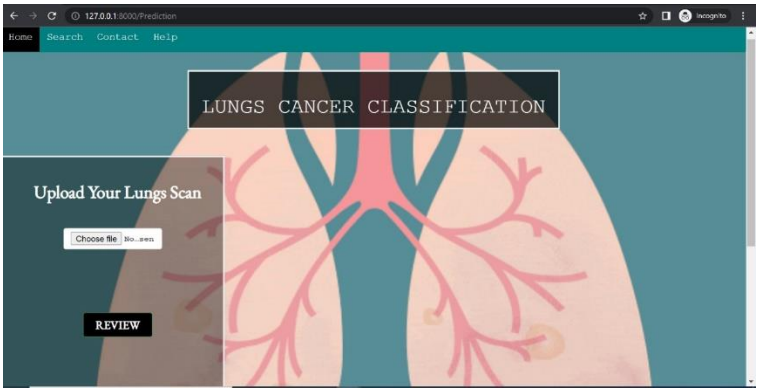
Our CNN model consists of six convolution layers of 32, 64, 64, 64,128,128 filters respectively and 6 max-pooling layers of size (2, 2) and then one hidden layer in fully connected layer with 64 node. Except the output layer, in all layers Relu activation used. As it is the multiclass classification problem, so softmax is used in output layer as activation function.

About Front End:

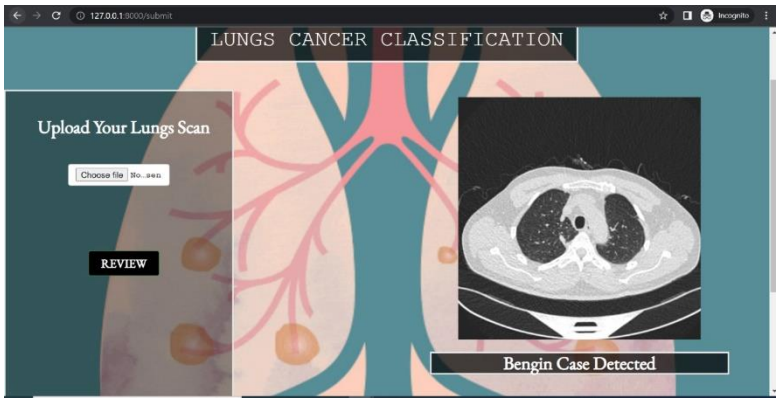
To build the front-end of the application, Flask was used to create an interactive web interface that allows users to upload an image of a lung and receive a prediction on whether the case is malignant, benign or normal.

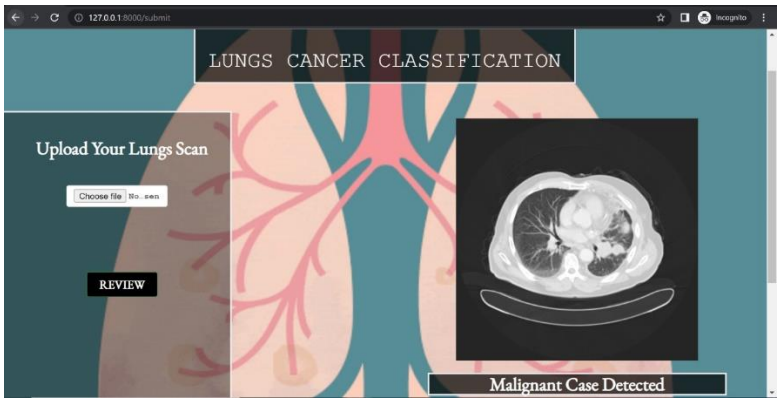
Main page of our website look like this as shown below:

After getting started, user will be redirected to the following page:



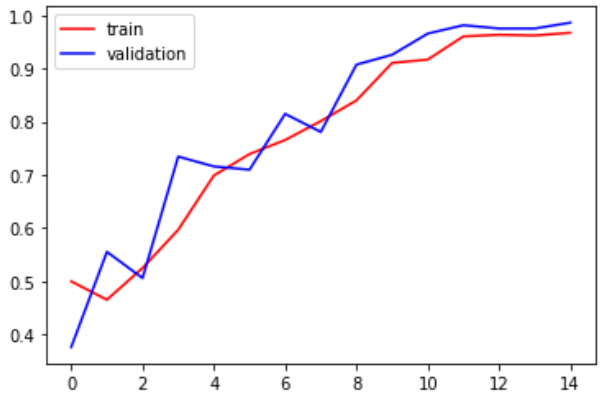
And classification interfaces for three classes are:



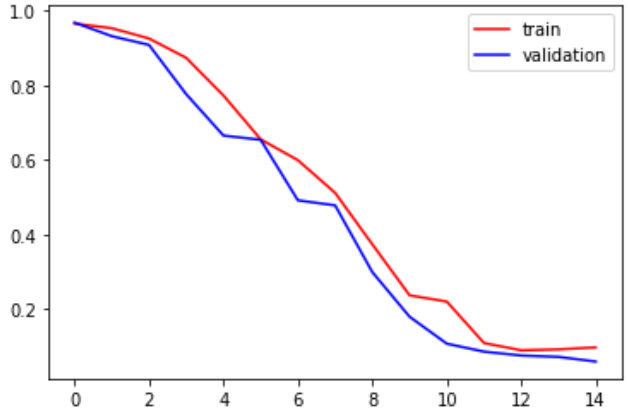


Result:

Till 15 epochs, we achieved accuracy of almost 98% without over fitting as shown in the given graph



Validation and training loss by using this CNN model, we achieved till 0.05



Conclusion:

In this project, a CNN model was developed in Python to classify lung cancer. The front-end of the application was built using Flask, a Python library for creating interactive web applications. The model achieved an accuracy of 98% on the test dataset and the Flask application was able to predict whether a lung cancer is “Benign case”, “Malignant case” or “Normal case” with high accuracy. This project demonstrates the effectiveness of using deep learning techniques such as CNNs in health field and the ease of building web applications in Python using libraries such as Flask.However, it's important to note that CNNs are not perfect and may not always give accurate results. Also, it's always recommended to consult with a radiologist for final diagnosis