Project Report

On

Application of Deep Learning Models for Lung Cancer Type Detection

Submitted

In partial fulfilment

For the award of the Degree of

PG-Diploma in Artificial Intelligence

C-DAC, ACTS (Pune)

Guided By: Mrs. Swapna E.

Submitted By:

Aditya Desai (220340128001)

Mayur Patil (220340128025)

Nikhil Rajesh Londhe (220340128032)

Piyush Nagpure (220340128034)

Prashant Kumar (220340128035)

Centre for Development of Advanced Computing (C-DAC), ACTS (Pune)

Acknowledgement

Thank you to the guide, Ms. Swapna E. for her constant guidance and helpful suggestions for preparing this project. We express our deep gratitude towards her for her inspiration, personal involvement, constructive criticism that she provided us along with technical guidance during the course of this project.

Thank you to the teachers at CDAC Pune, the co-ordinators, Dr. Mrs. Priyanka Ranade, Ms. Seema, Ms. Gayle Fernandes, the process owner, Mrs. Namrata Ailawar, the program head, Mrs. Risha P R, the head of the department, Mr. Gaur Sunder, the team at CDAC Pune who provided us this opportunity to carry this prestigious project and enhance our learning in various technical fields.

Thank you to our parents and families, and to the ones because of whom this project has been developed.

Thank you to the classmates for maintaining a co-operative, collaborative, and joyous environment.

Thank you to CDAC.

Aditya Desai (220340128001)

Mayur Patil (220340128025)

Nikhil Rajesh Londhe (220340128032)

Piyush Nagpure (220340128034)

Prashant Kumar (220340128035)

Abstract

Cancer is the uncontrollable cell division of abnormal cells inside the human body, which can spread to other body organs. It is one of the non-communicable diseases (NCDs) and NCDs accounts for 71% of total deaths worldwide whereas lung cancer is the second most diagnosed cancer after female breast cancer. Cancer survival rate of lung cancer is only 19%. There are various methods for the diagnosis of lung cancer, such as X-ray, CT scan, PET-CT scan, bronchoscopy and biopsy. However, to know the subtype of lung cancer based on the tissue type H and E staining is widely used, where the staining is done on the tissue aspirated from a biopsy. Studies have reported that the type of histology is associated with prognosis and treatment in lung cancer. Therefore, early and accurate detection of lung cancer histology is an urgent need and as its treatment is dependent on the type of histology, molecular profile and stage of the disease, it is most essential to analyse the histopathology images of lung cancer.

Hence, to speed up the vital process of diagnosis of lung cancer and reduce the burden on pathologists, Deep learning techniques are used. These techniques have shown improved efficacy in the analysis of histopathology slides of cancer. Several studies reported the importance of convolution neural networks (CNN) in the classification of histopathological pictures of various cancer types such as brain, skin, breast, lung, colorectal cancer. In this study tri-category classification of lung cancer images (normal, adenocarcinoma and squamous cell carcinoma) are carried out by using ResNet 50, VGG-16, VGG-19, Inception_ResNet_V2 and DenseNet for the feature extraction and triplet loss to guide the CNN such that it increases inter-cluster distance and reduces intra-cluster distance.

Keywords: ResNet 50, CNN, VGG-16, VGG-19, Inception_ResNet_V2 and DenseNet, Histopathology Images

Table of Contents

Introduction

Introduction

Lung cancer type detection is of significance in the world where respiratory health is being discussed across geographies – in general - and during pandemic – in specific. Importance of lung functioning has been in the spotlight for much time. There are external factors, such as environment, climate change, that might impact the respiratory health of people, nevertheless, the people who have been diagnosed with the disease would be helped with the detection of the sub-type.

A number of people succumb to lung cancer every year. A type of lung cancer is "non-small cell". LUAD, LUSC, large cell are the sub-types; LUAD translates to Adenocarcinoma, and LUSC to Squamous cell carcinoma. Lung biopsies are typically used to diagnose sub-type and stage. Therapies are applied depending on the sub-type, stage, and mutation.

Microscopic images can be of size 10000 to 100000 pixels. These images are useful for the detection of the sub-type.

General idea about our model:

The detection of cancerous cells can be done using images of slides. The slides are pre-processed. Then the deep learning models are trained on the slide images .The trained model is then validated using test data. This is expected to help pathologists classify whether lung cancer is present or not.

Histopathology-An introduction:

Histopathology refers to the study of the signs of the disease using the microscopic examination of a biopsy or surgical specimen that is processed and fixed onto glass slides.

Stages of the preparation process of the tissue slides are as follows-

1.Fixation: Biological tissues are fixed with chemical fixation using formaldehyde or glutaraldehyde solution to protect the cells. This aims to prevent tissue autolysis and putrefaction.

2.Processing: It involves dehydration and clearing. Dehydration is used to extract water from the gross tissue and substitute it with a certain concentration of alcohol which solidifies and helps incise superfine sections of the specimen. Clearing removes the dehydrator with a material that will be the solvent in both the embedding paraffin and the dehydrating agent.

3.Tissue Embedding: Tissues are carefully positioned in a medium such as wax, so that it will provide enough external support to allow very thin sectioning on solidification.

4.Sectioning: This generates sufficient superfine slices of tissue samples such that the details of the microstructure characterization of the cells can be noticed using microscopy methods.

5.Staining: The final step in preparing tissue for light microscopy is to stain it and mount it on the slide.

Imaging of slides:

The different ways to image the tissues are-

1.The tissue is dyed with different stains for microscope visualisation. Pathologists generally use Hematoxylin-Eosin (H&E). Hematoxylin stains cell nuclei blue, while Eosin provides pink colour to cytoplasm and connective tissue. the nuclei images from the H&E images can be computed using decomposition techniques or using methods that utilize the fact that blue wavelengths are absorbed less than green and red channels by the Hematoxylin dye.

2.Immuno-fluorescence imaging-It uses molecular markers based on chromogenic dyes (such as DAB), or fluorescent dyes (such as Cy dyes or Alexa dyes). In fluorescence the excited molecule gives us some of its vibrational energy in collision with other molecules, so that the downward radiative transition originates from a lower vibrational level in upper electronic state. Here, DAPI is used to stain nuclei blue.

3.Spectroscopic Imaging- It combines the strengths of the method of point spectroscopy and spectral imaging, building spatial imaging of the human tissue in a multitude of wavelength regimes. Point spectroscopy uses the principle of Raman scattering. Raman Scattering refers to the form of scattering in which a monochromatic radiation or a radiation of very narrow frequency band when scattered consists of not only of the radiation of incident frequency but also of above and below frequency than that of incident beam's frequency. Spectral imaging is carried out by building an image cube with slices corresponding to images of the same scene obtained by incident light at differing wavelengths.

The Whole slide Images generated by the above processes are studied by pathologists to arrive at a diagnosis.

Application of machine learning to histopathological studies:

Whole slide images (WSI) are mapped to one of the disease categories for diagnostic purpose using different machine learning and deep learning methods.

This offers following advantages-

1.Improvement of classification accuracy can lead to reduced variability in interpretations and prevent overlooking by investigating all pixels within WSIs.

2.Segmentation of Region of interest.

3.Content Based Image Retrieval (CBIR) retrieves images similar to a query image. In digital pathology, CBIR systems are useful in many situations, particularly in diagnosis, education, and research. For example, CBIR systems can be used for educational purposes by students to retrieve histopathological images of tissues. In addition, such systems can also be helpful to professional pathologists, particularly when diagnosing rare cases.

4.By analysing the data obtained, new clinicopathological relationships can be developed using computational methods.

Problems with Histopathological Image Analysis:

1.Very large Image size-It leads increase in the number of parameters to be estimated which increases computational power and memory requirement.

2.Insufficient labelled images

3. Different Levels of Magnification Result in Different Levels of Information - Information regarding cell shape is well captured in high-power field microscopic images, but structural information such as a glandular structure made of many cells are better captured in a lower-power field. Because cancerous tissues have both cellular and structural atypia, images taken at multiple magnifications would each contain important information. Pathologists diagnose diseases by acquiring different kinds of information from the cellular level to the tissue level by changing magnifications of a microscope.

4. WSIs are created through multiple processes. At each step undesirable effects, which are unrelated to the underlying biological factors, could be introduced.

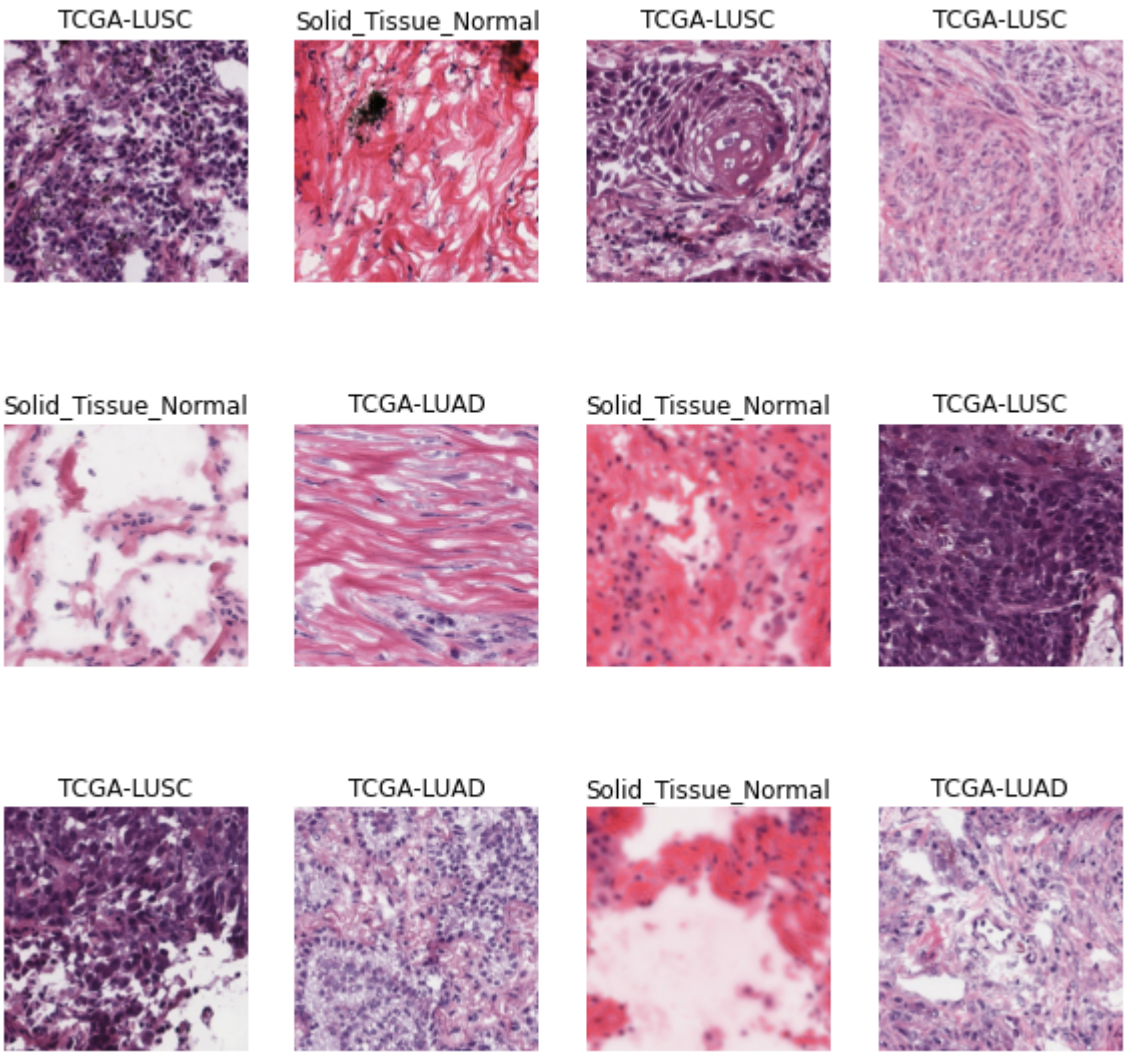
Literature Survey

The two common subtypes of lung cancer were focused on in the paper titled *"Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning"* by Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, Aristotelis Tsirigos. LUAD and LUSC are the two sub-types. Two classification processes were done, the first where the images were classified between 'normal lung tissue' and 'affected', and, the second where the images were classified between 'normal lung tissue', 'LUAD', and 'LUSC'. The model used was "inception v3". The data obtained was from TCGA – The Cancer Genome Atlas.

The images used are stained images of histopathology. Two approaches, such as Transfer Learning and Fully-trained Inception architecture, were used. The images were of varying pixels. Training, validation, and testing split was 70, 15, 15 respectively. The pixels were made 512 * 512.

Yu et al. have been mentioned to have studied the use of models such as Random Forest, SVM - which is Support Vector Machines -, Naïve Bayes. The results there have been different from the Inception v3 model.

Data Visualisation



Implementation

Implementation

1. Use of Python Platform for writing the code with TensorFlow, Flask, Streamlit.
2. Hardware and Software Configuration:

   Hardware Configuration:

- CPU: 8 GB RAM, Quad core processor
- GPU: 25 GB RAM, T4(Colab Pro)

  Software Required:

- Anaconda: It is a package management software with free and open-source distribution of the Python and R programming language for scientific computations (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify deployment.
- Colab Pro
- Spyder: Spyder, the Scientific Python Development Environment, is a free integrated development environment (IDE) and open-source scientific environment that is included with Anaconda written in Python, for Python, and designed by and for scientists, engineers and data analysts. It includes editing, interactive testing, debugging, and introspection features with the data exploration, interactive execution, deep inspection, and beautiful visualisation capabilities of a scientific package.

Data Gathering

The overall algorithm that the process follows is-

1. Data is downloaded from gdc portal by following steps-

   download the client from
   https://gdc.cancer.gov/access-data/gdc-data-transfer-tool

   Create and download a manifest and metadata json file from the gdc website for the whole slides images of interest

   Download images using the manifest and the API: gdc-client.exe download -m gdc_manifest.txt

2. The svs images are tiled and converted in jpg
3. The jpg images are sorted into train, test and validation set and put into appropriate classes
4. Training and validation is run.

Methodology

The methodology that shall be followed is as follows.

- Obtaining the dataset
- Pre-processing, if any
- Training, validation, testing split
- Model selection
- Model training
- Classification
- Result

Obtaining the dataset:

The images shall be in the JPEG format.

Pre-processing, if any:

The images shall be of the size 512 * 512.

Training, validation, testing split:

The training, validation and testing split can be made as mentioned in the sessions as 80(as per 80, 20), 10, 10. There shall be three folders as with the names mentioned above. Approximately the training set would contain 100,000 files, and the testing and validation 10,000 each, on an average, for each class. Update: 9000 files in training set, 1200 in test and validation each, approximately, for 3 classes. Second update: For model 1.1, which is the model which used 25 percent of the dataset, there were 19895 files in the training set, and around 2400 files in the validation and test sets each, for 3 classes. Third update: Around 9000 files in the training dataset, 12000 in the overall dataset.

Model selection:

VGG, DenseNet, Convolutional Neural Network, ResNet, Inception

Model Description:

The models we used are:

Visual Geometry Group:

There are three variants of the VGG model, first, the VGG 16 layers, second, the VGG 19 layers, and third, the fusion. The 19-layer has shown to give better performance, hence the model to be used is this one. The 16-layer model has been used later as well. VGG model which is available in tensorflow module.

Convolutional Neural Networks (CNN):

A convolutional neural network is a special type of deep learning model which performs extremely well for image classification purposes. A CNN basically consists of an input layer, an output layer and a hidden layer which can have multiple numbers of layers. A convolution operation is

performed on these layers using a filter that performs 2D matrix multiplication on the layer and filter.

The different layers of CNN model consists of the following:

Convolutional function

MaxPooling2D function

Dropout function which drops random nodes, s

Flatten function

In all the layers, a Relu activation function is used except the output layer in which we have used Softmax. Adam Optimiser alongwith Reduce on plateau function is used to better train the model.

Model training:

1. Dataset Loading:

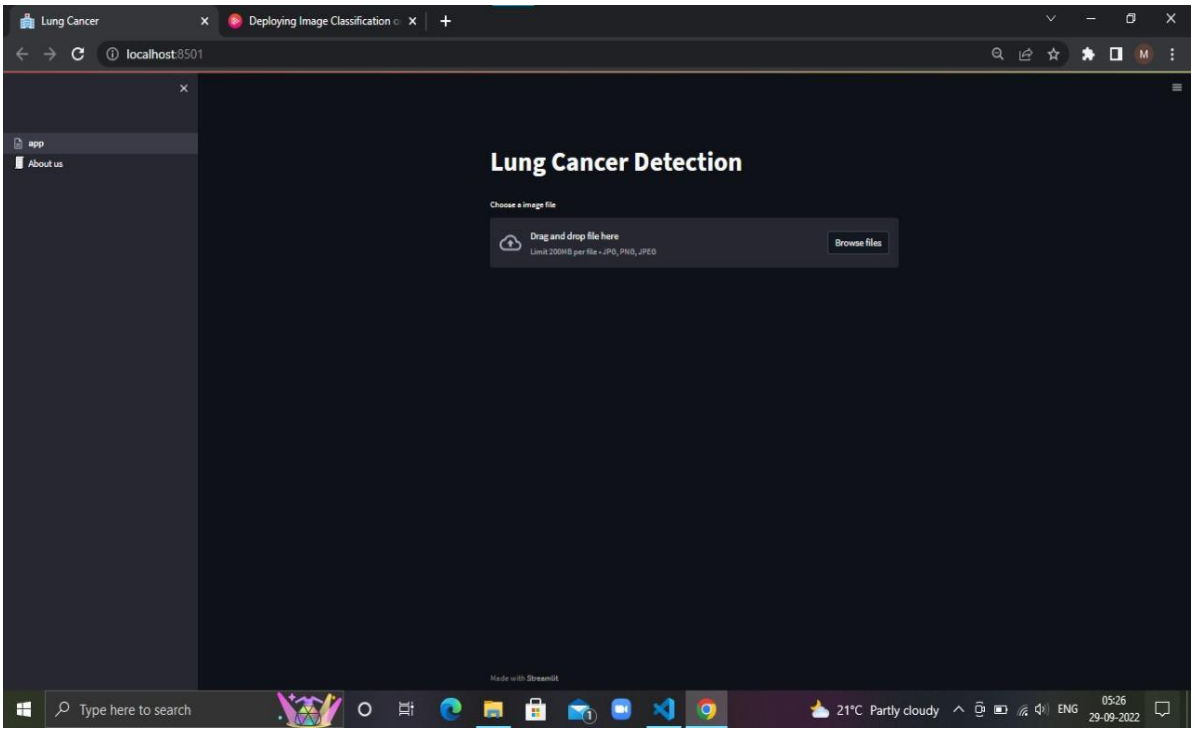
   Dataset size: Over 30 GB; For model 1, which is the VGG model which used 20 percent of the dataset, the size was around 6 GB. For model 1.1, the size was around 7.5 GB.

2. TensorFlow over PyTorch
3. Weights: imagenet
4. Classes: 3
5. Image size: 512 * 512/224 * 224/112 * 112/256 * 256 were the alternatives, from which 256 * 256 was selected.
6. Epochs: First, 10 were taken. Then, 50.

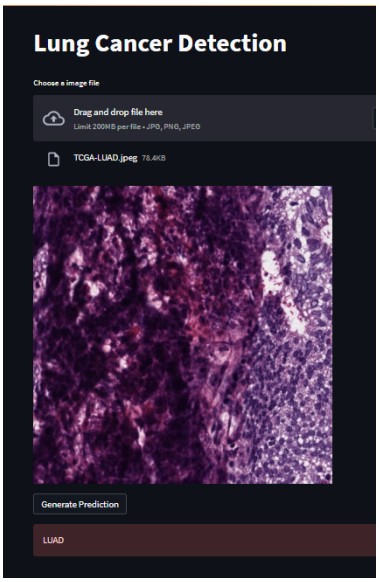
Classification:

Multi class classification with three classes namely, Solid_Tissue_Normal, TCGA-LUAD and TCGA-LUSC.

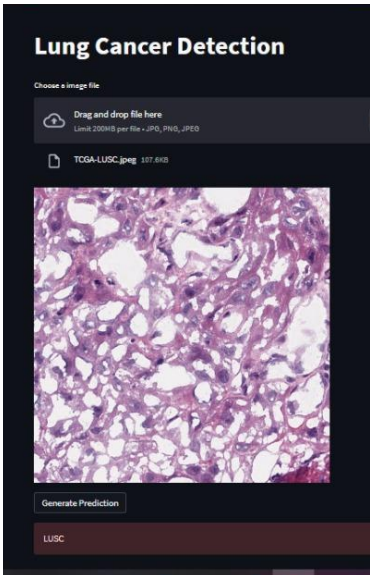
Deployment:



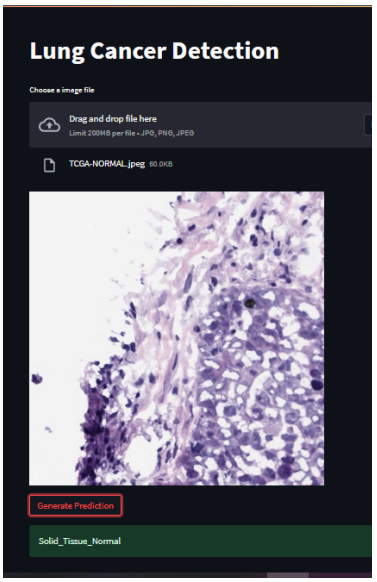
Homepage



LUAD



LUSC



Normal

Result:

Name of the model	Training accuracy (in %)	Training loss	Test accuracy (in %)	Test loss	Validation accuracy (in %)	Validation loss
VGG19	86.72	0.338	85.47	0.3489	85.60	0.356
Inception	70.1	1.19	69.45	1.086	70.39	1.138
ResNet	97.41	0.082	95.86	0.1203	96.22	0.109
DenseNet	90.97	0.231	91.41	0.2153	91.04	0.2533
VGG16	88.49	0.2889	89.45	0.2711	89.06	0.2776

Results

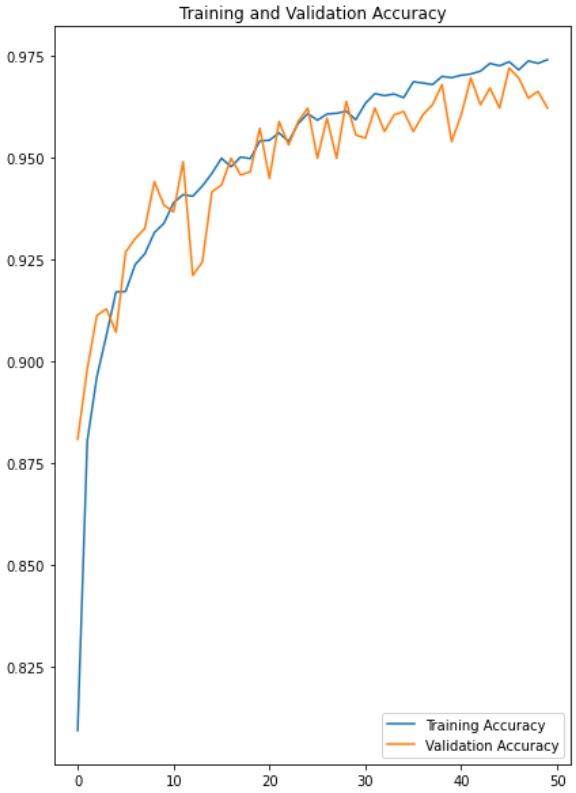
Results Epochs:

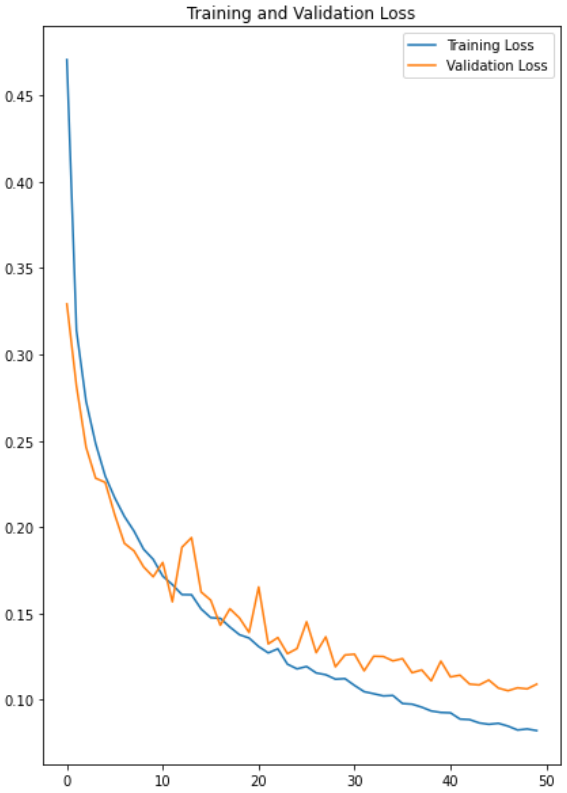
```
Epoch 45/50
311/311 [==============================] - 24s 76ms/step - loss: 0.0858 - accuracy: 0.9726 - val_loss: 0.1114 - val_accuracy: 0.9622
Epoch 46/50
311/311 [==============================] - 24s 76ms/step - loss: 0.0864 - accuracy: 0.9736 - val_loss: 0.1068 - val_accuracy: 0.9720
Epoch 47/50
311/311 [==============================] - 24s 76ms/step - loss: 0.0848 - accuracy: 0.9716 - val_loss: 0.1052 - val_accuracy: 0.9696
Epoch 48/50
311/311 [==============================] - 24s 76ms/step - loss: 0.0825 - accuracy: 0.9738 - val_loss: 0.1069 - val_accuracy: 0.9646
Epoch 49/50
311/311 [==============================] - 24s 77ms/step - loss: 0.0831 - accuracy: 0.9732 - val_loss: 0.1063 - val_accuracy: 0.9663
Epoch 50/50
311/311 [==============================] - 24s 77ms/step - loss: 0.0822 - accuracy: 0.9741 - val_loss: 0.1090 - val_accuracy: 0.9622
```

- We have run 50 epochs for this model
- We were able to achieve an accuracy of 97% on the Train dataset and 96% on validation dataset.
- The training and Validation Losses were comparable.

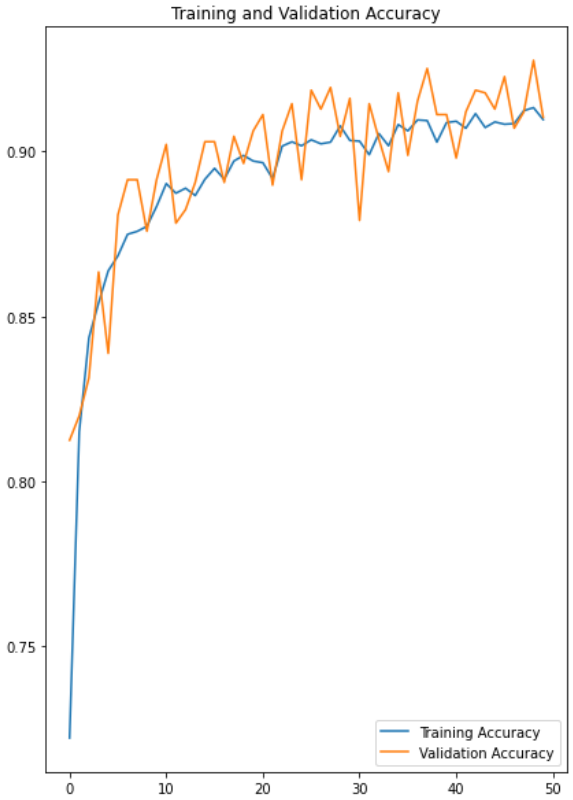
Output Graphs of accuracy and loss for training and validation

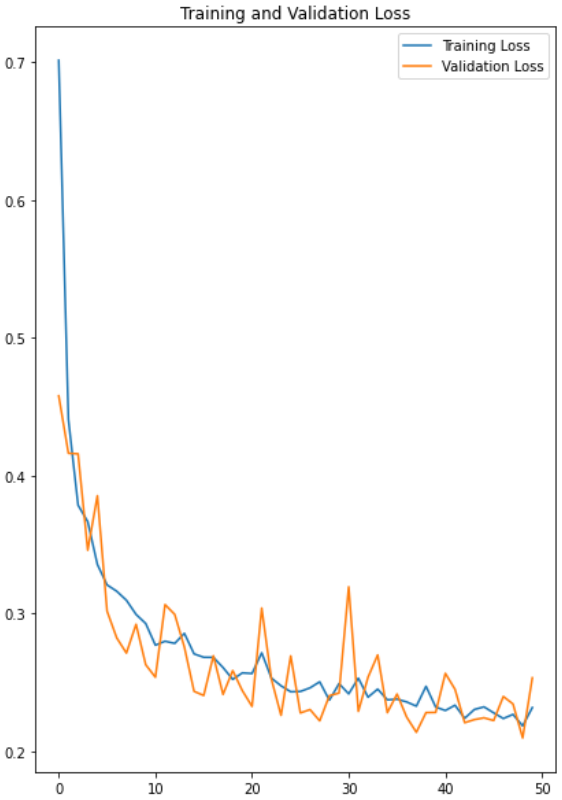
- ResNet Model-



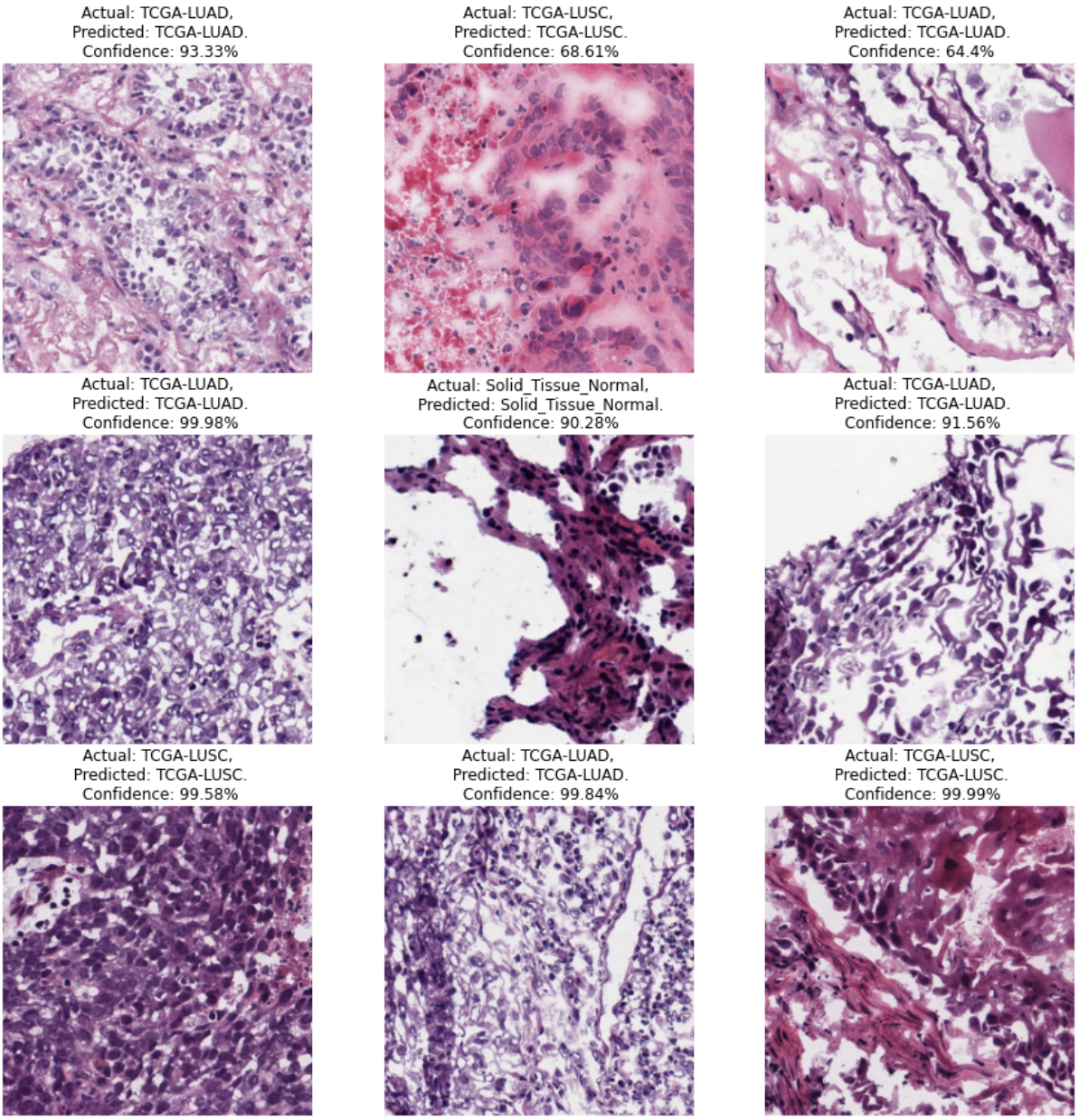


- DenseNet Model-



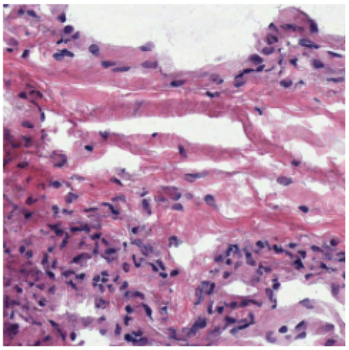


Predicted result for DenseNet Model:

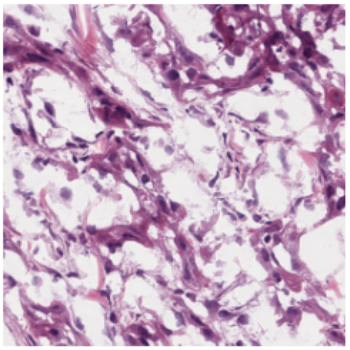


Predicted result for ResNet Model:

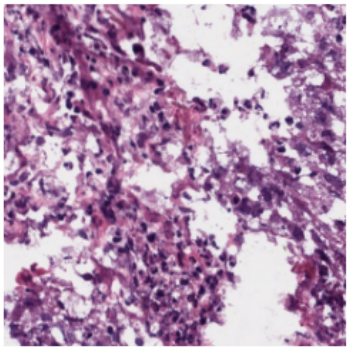
Actual: Solid_Tissue_Normal,
Predicted: Solid_Tissue_Normal.
Confidence: 99.9%



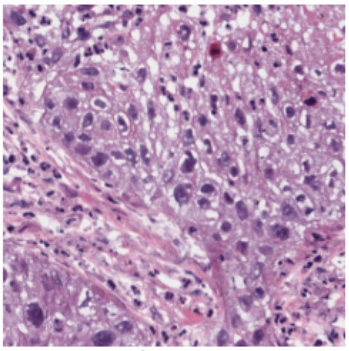
Actual: TCGA-LUSC,
Predicted: TCGA-LUSC.
Confidence: 98.8%



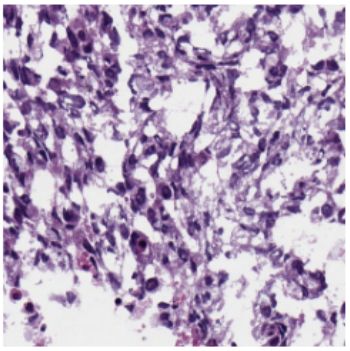
Actual: TCGA-LUSC,
Predicted: TCGA-LUSC.
Confidence: 68.75%



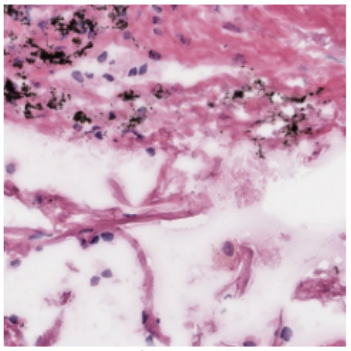
Actual: TCGA-LUSC,
Predicted: TCGA-LUSC.
Confidence: 99.06%



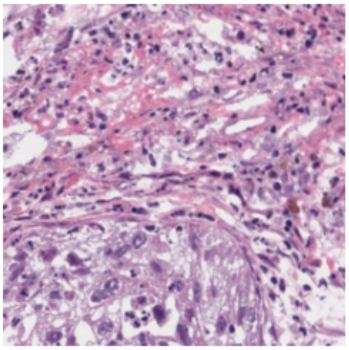
Actual: TCGA-LUAD,
Predicted: TCGA-LUAD.
Confidence: 100.0%



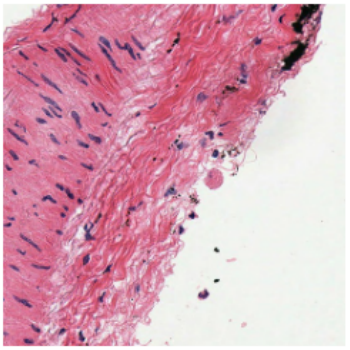
Actual: Solid_Tissue_Normal,
Predicted: Solid_Tissue_Normal.
Confidence: 99.65%



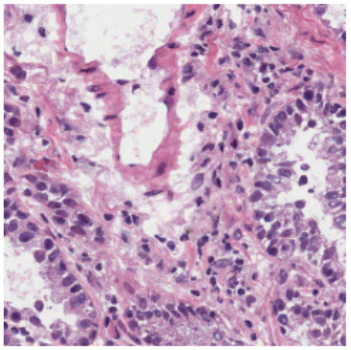
Actual: TCGA-LUSC,
Predicted: TCGA-LUSC.
Confidence: 85.98%



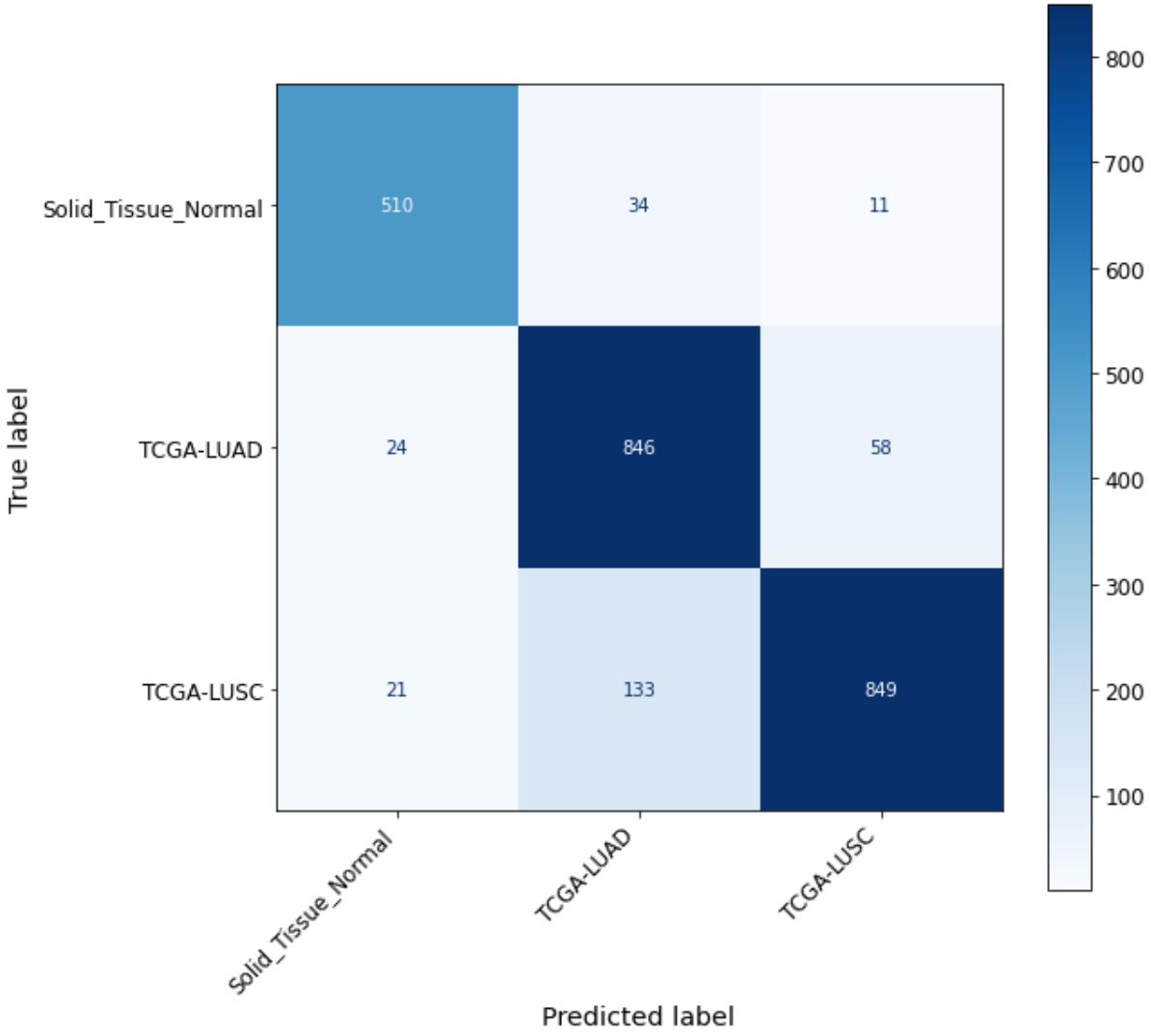
Actual: Solid_Tissue_Normal,
Predicted: Solid_Tissue_Normal.
Confidence: 99.33%



Actual: TCGA-LUAD,
Predicted: TCGA-LUAD.
Confidence: 69.09%



Confusion Matrix-



Conclusion

Conclusion

In this Python project, we have tried to train a deep learning model on cancer slide images. We used TensorFlow to detect images, and then we used different CNN models to classify the images.

- Lung Cancer is one of the most common forms of cancer present in modern times. Pathologists have to employ a lot of resources to detect cancerous cells.
- However, if the process of detecting cancerous cells can be automated using deep learning models it could help in early detection and save time for pathologists.
- We have trained Deep learning CNN, VGG, DenseNet, ResNet, Inception models. We have tweaked the hyperparameters ,activation functions in different layers to achieve decent accuracy.
- The model can be adapted for scalability, better convergence, and better accuracy.

Acronyms

The explanation of acronyms used is mentioned at the respective areas where the acronym was used, as much as suitable. Further explanations are given as follows.

- VGG: Visual Geometry Group
- TCGA: The Cancer Genome Atlas
- CNN: Convolutional Neural Network

Reference

For reference, the GitHub page that was made available, from where the information about the dataset and the model used, Inception v3, was obtained:

https://github.com/ncoudray/DeepPATH

For the information of the Visual Geometry Group models:

Visual Geometry Group - University of Oxford

For the VGG, ResNet, DenseNet, Inception Architecture of TensorFlow:

https://www.tensorflow.org/api_docs/python/tf/keras/applications/vgg19/VGG19