UE20CS461A – Capstone Project Phase – 2

Project Progress Review #1

Project Title : **From Diagnosis to Remission: Understanding Cancer Detection and Staging.**

Project ID : 103

Project Guide : Dr. Prema R

Project Team :

Akshat Bhandari	:	PES2UG20CS030
Ayush Singh	:	PES2UG20CS080
Ayan Aggarwal	:	PES2UG20CS079
Ankur Kumar Dubey	:	PES2UG20CS054

Outline

- Abstract and Scope of the Project
- Capstone Project Phase – 1
  - Summary of work
  - Inferences drawn from Literature Survey
- Expected Deliverables
- Gantt chart

Abstract

Problem Statement:

- Lung cancer is a significant health issue worldwide, with early detection and diagnosis being crucial for successful treatment and improved patient outcomes. However, interpreting medical images manually can be time-consuming, error-prone, and often requires specialized knowledge. Therefore, there is a need for a reliable and automated system to assist in the early detection and diagnosis of lung cancer and to provide valuable information to medical professionals.

Introduction:

- The lung cancer detection website is a web-based application designed to assist medical professionals in the early detection and diagnosis of lung cancer. The website leverages machine learning algorithms, image processing techniques, and specialized knowledge to analyze medical images and provide a diagnosis based on the stage of cancer present.
- The website's main functionality will be divided into three parts. The first part will be responsible for uploading medical images, which will be processed and analyzed by the machine learning model. The second part will display the results of the analysis in an easily understandable format, highlighting the areas of the image that are potentially cancerous and indicating the stage of cancer present. The third part of the website will include a reports section, providing additional analysis and information for medical professionals.

Abstract

- The reports section will include analysis covering treatment plans available, specialized hospitals for such treatment plans, and the patient's life expectancy based on the stage of cancer present. This information will be available only to radiologists and doctors, providing valuable insights into the best possible treatment options for the patient.

- The website's success will be measured by its ability to accurately detect the stage of cancer present in medical images, assist medical professionals in providing the best possible treatment options for patients, and reduce the time required for manual interpretation of medical images.

- In summary, the lung cancer detection website will be a valuable tool for medical professionals, allowing them to quickly and accurately detect and diagnose lung cancer and provide patients with the best possible treatment options. By leveraging machine learning algorithms, image processing techniques, and specialized knowledge, the website will provide a reliable and automated system for the early detection and diagnosis of lung cancer.

Suggestions

Suggestions

- Implement a color-coding scheme to assist Medical Professionals/Patients in identifying abnormal/cancerous regions based on cancer stage.
- Create a reports section to provide medical professionals with in-depth analysis and insights, such as treatment plans available, specialized hospitals, and life expectancy.
- Ensure the website's interface is user-friendly and intuitive, allowing medical professionals to quickly and easily upload and analyze medical images.
- Focus on accuracy and reliability in the project's development to provide medical professionals with a valuable tool for early detection and diagnosis of lung cancer.

Summary of Work Done in Capstone Project Phase - 1

Completed phases/Work summary so far :

Feasibility study: In Phase 1, we assessed the project's feasibility, confirming its practicality within resource constraints.

Literature survey: This phase involved an extensive literature review, highlighting key findings and trends in the field to inform our project direction.

Model and Algorithm shortlisting: We selected potential models and algorithms suitable for our project, laying the foundation for further development.

High-level design and framework: During Phase 1, we created a high-level architectural design, outlining the project's structure and technological framework.

Idea of the Workflow: We conceptualized the project workflow, mapping out the sequential steps and processes that will be followed throughout the project's development.

Expected Deliverables

1. List of tasks/Modules to be elaborated in discussion with the guide.

1. For example, If your project consist of Data Preprocessing, the following should be explained,
   a) Data Collection
   b) Data Preparation
   c) Data Input
   d) Data Pre-processing
   e) Data Visualization
   f) Data Interpretation
   g) Storage

Expected Deliverables

3. Tabulate the individual contribution of the team members with the following,
    a) Tasks/Modules assigned
    b) Development (no. of lines of code & time spent)

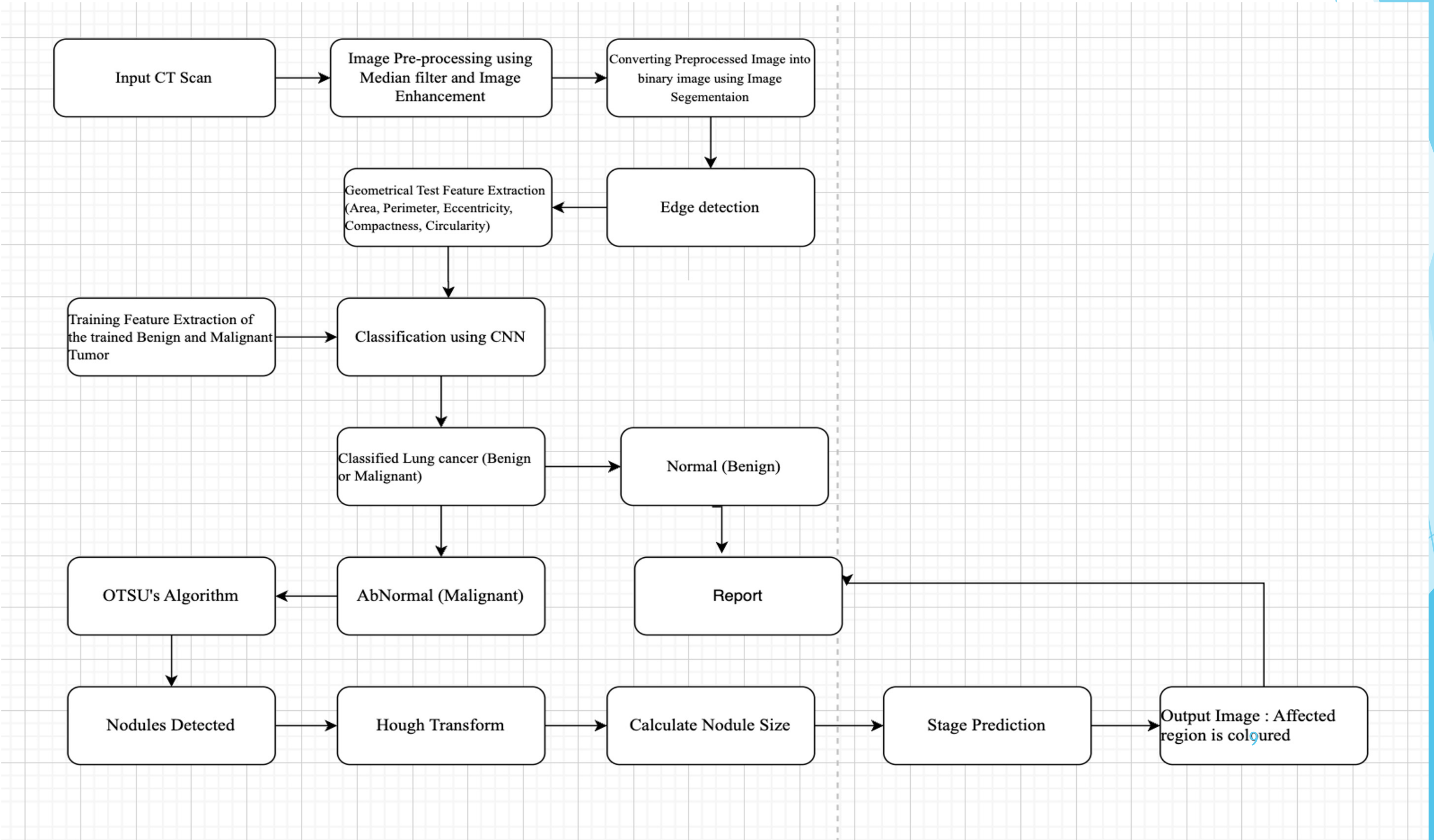
4. List the SDK / API / Model / Jar/ DLL / Tools / Technologies used – Open-Source/ Licensed.

5. Testing for the module that is completed.

6. Demonstration and Result of modules completed.

7. Tabulate the timeline for all the tasks/modules.

Architecture



Input CT Scan

Image Pre-processing using Median filter and Image Enhancement

Converting Preprocessed Image into binary image using Image Segementaion

Edge detection

Geometrical Test Feature Extraction (Area, Perimeter, Eccentricity, Compactness, Circularity)

Training Feature Extraction of the trained Benign and Malignant Tumor

Classification using CNN

Classified Lung cancer (Benign or Malignant)

Normal (Benign)

Report

AbNormal (Malignant)

OTSU's Algorithm

Nodules Detected

Hough Transform

Calculate Nodule Size

Stage Prediction

Output Image : Affected region is coloured

Expected Modules

There are 2 main modules of our project:

Module 1:

1) To make the images more clear , we apply some filters on CT scan like binary filtering .
2) After obtaining a clear image our model will check whether the cancer is malignant or benign .
3) If the cancer is benign we can tell that the cancer is non cancerous and and we can build a report for the same.

Module 2:

1) If the cancer is malignant detect the largest affected nodule.
2) On the basis of dimensions of largest affected nodule detect the stage of the cancer.

About Data

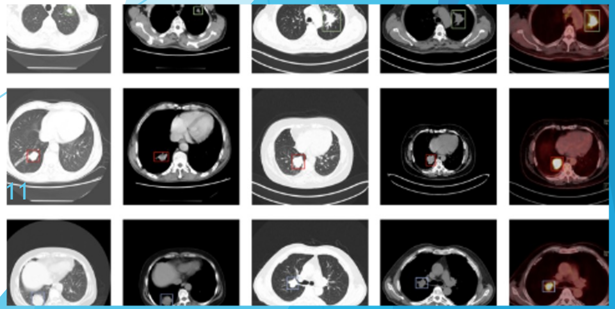
The "IQ-OTH/NCCD - Lung Cancer Dataset" is a medical dataset collected from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD). It was gathered over a period of three months in the fall of 2019 and contains CT scans of individuals diagnosed with lung cancer at different stages, as well as scans from healthy subjects. Below is a description of the dataset:

Description:

- **Data Source:** The dataset was collected at the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) by specialists, oncologists, and radiologists.

Dataset Contents:

- **Total Images:** The dataset comprises a total of 1,190 images representing CT scan slices from 110 cases.
  - **Classes:** The cases are categorized into three classes: normal, benign, and malignant.
  - Malignant Cases: 40 cases
  - Benign Cases: 15 cases
  - Normal Cases: 55 cases



Important Libraries

Here are a few significant libraries :

- **numpy**: Fundamental library for scientific computing.
- **pandas**:Data manipulation and analysis, especially with tabular data.
- **matplotlib**: Data visualization.
- **PIL (Python Imaging Library)**:Opening, manipulating, and saving images.
- **seaborn**: High-level data visualization.
- **cv2 (OpenCV)**: Computer vision and image processing.
- **os**:Operating system-dependent functionality (e.g., file operations).
- **sklearn (scikit-learn)**:Machine learning for data mining and analysis.
- **imblearn**:Library for dealing with imbalanced datasets.
- **tensorflow**:Deep learning framework.
- **keras**:High-level deep learning API.

Data Preprocessing steps

Setup of Data Directories:

In the directory variable, we specified the directory path where the dataset is stored.

Category Types:

'Benign cases,' 'Malignant cases,' and 'Normal cases' are the three class categories we defined in the categories list.

Image Size and Loading Analysis:

To create a dictionary that counts the occurrences of various image sizes for each class, we iterated through the directories of each class, loaded the images, and examined their sizes.

Image Data Extraction and Normalization:

We loaded the images, resized them to a common size (256x256 pixels), and converted them to grayscale.

Applying filters

We applied various filters including median filter to remove salt and pepper noise.

Data Preprocessing steps

Data Splitting:

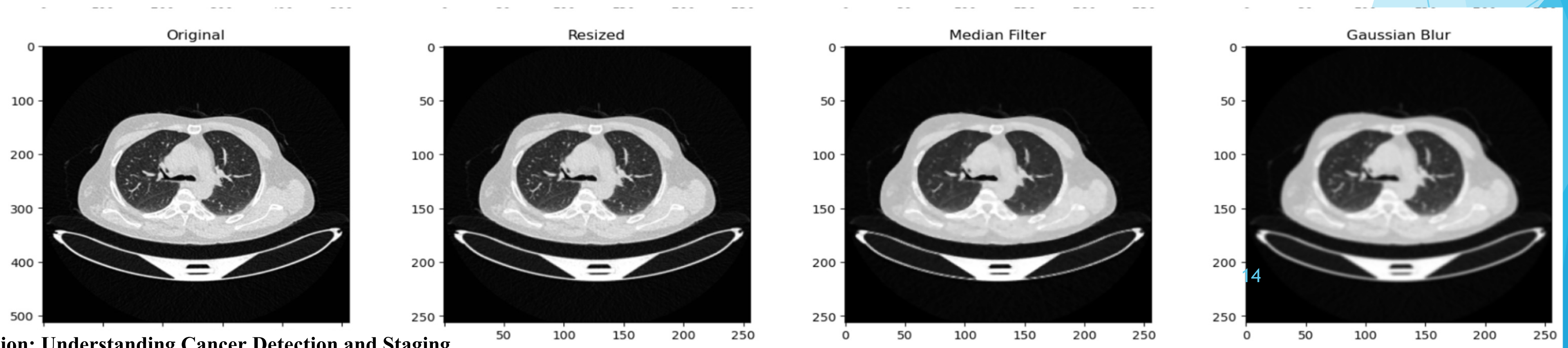
We split the data into training and validation sets using train_test_split, ensuring that the class distribution is maintained in both sets.

Data Balancing (SMOTE):(Synthetic Minority Over-sampling Technique)

We applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distribution in the training data. This was done to address the imbalance between benign, malignant, and normal cases.

Data Reshaping After SMOTE:

After SMOTE, we reshaped the training data to include the single-channel dimension



Module 1

In order to move on to stage module, module one involves identifying benign, malignant, and normal cases. We used the CNN model to divide this classification into three classes.

There are two layers in our CNN model.

- The first layer is a Convolutional layer with 64 filters. It takes input data with a shape corresponding to images with height, width, and channels. The activation function applied to this layer is ReLU (Rectified Linear Unit), and it's followed by a Max Pooling layer.

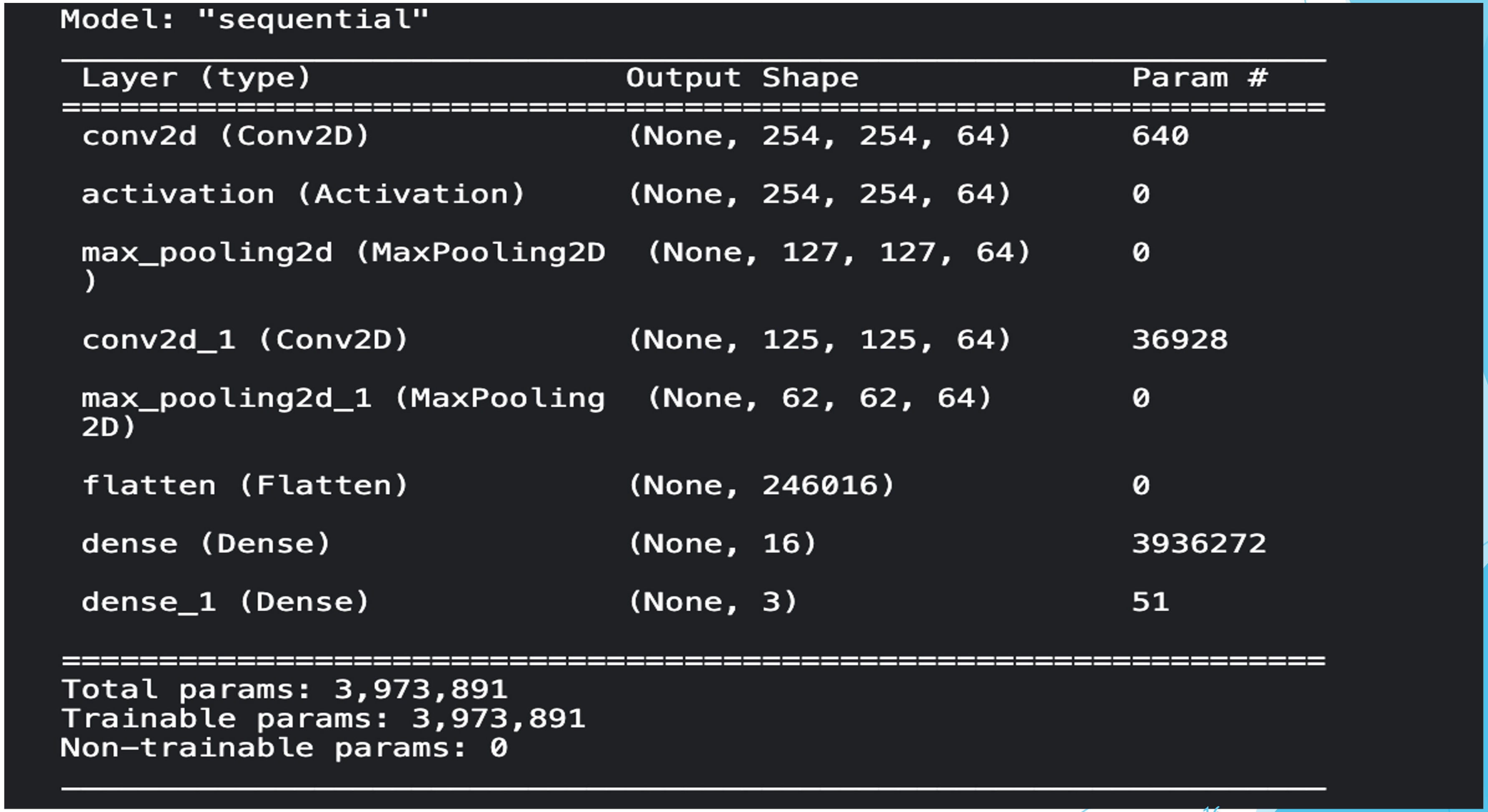
- The second layer is also a Convolutional layer with 64 filters , and it utilizes ReLU activation. Following this convolutional layer, there is another Max Pooling layer.

- After these convolutional and pooling layers, the network incorporates a Flatten layer.

- Next, there is a Dense (fully connected) layer comprising 16 neurons.

- Finally, in the output layer, there are 3 neurons corresponding to the different classes, and it uses the softmax activation function, which is commonly used for multi-class classification problems.

CNN Model Overview



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 254, 254, 64)      640

 activation (Activation)     (None, 254, 254, 64)      0

 max_pooling2d (MaxPooling2D  (None, 127, 127, 64)     0
 )

 conv2d_1 (Conv2D)           (None, 125, 125, 64)      36928

 max_pooling2d_1 (MaxPooling  (None, 62, 62, 64)       0
 2D)

 flatten (Flatten)           (None, 246016)            0

 dense (Dense)               (None, 16)                3936272

 dense_1 (Dense)             (None, 3)                 51

=================================================================
Total params: 3,973,891
Trainable params: 3,973,891
Non-trainable params: 0
_________________________________________________________________
```

Model 1: Addressing Class Imbalance with SMOTE

Challenge

Our lung cancer dataset had an imbalance in the number of instances across the three classes - benign, malignant, and normal.

Solution

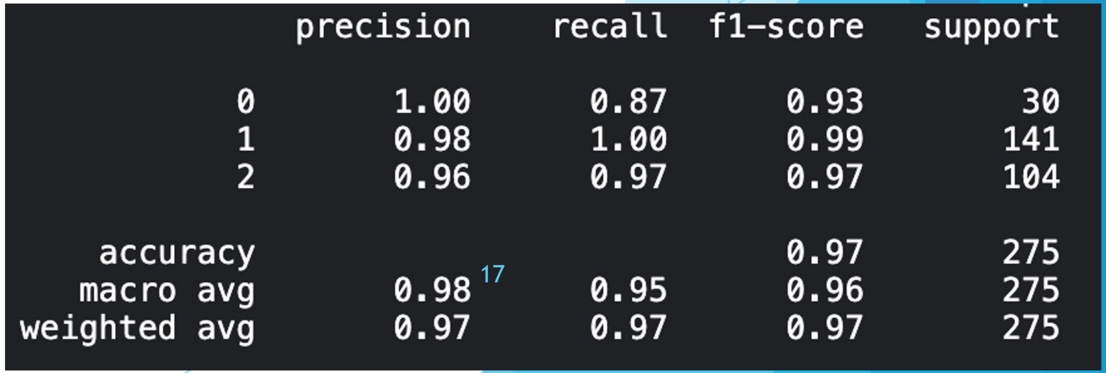
To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority classes, essentially balancing the dataset.

Outcome

This balanced dataset was used to train our Convolutional Neural Network (CNN) model.

Performance

The model achieved an accuracy of 94% on the validation data.



```
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        30
           1       0.98      1.00      0.99       141
           2       0.96      0.97      0.97       104

    accuracy                           0.97       275
   macro avg       0.98      0.95      0.96       275
weighted avg       0.97      0.97      0.97       275
```

Model 2: Class-Weighted Training

Challenge

Class imbalance remained a concern, but in this approach, we explored an alternative strategy without oversampling or undersampling.

Solution

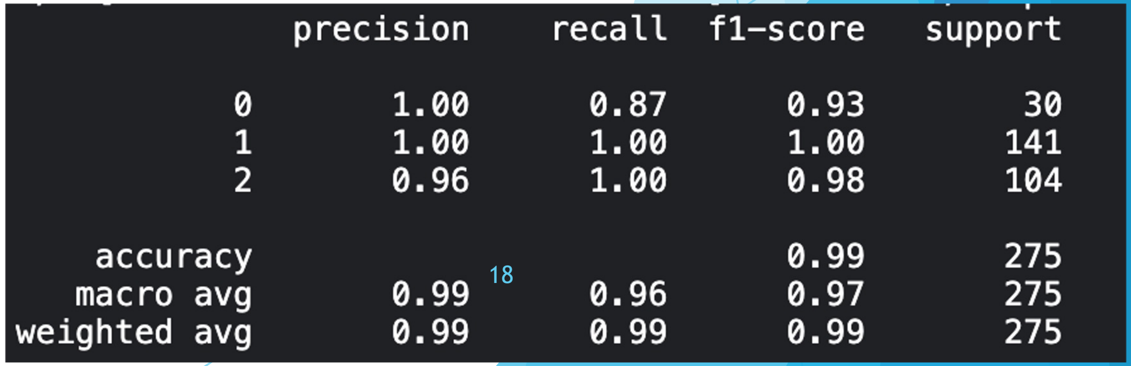
We calculated class weights based on the class distribution in the training data. These weights were incorporated during model training, giving higher importance to the minority classes.

Outcome

The model was trained with an emphasis on addressing class imbalance.

Performance

The model achieved an accuracy of 99% on the validation data.



```
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        30
           1       1.00      1.00      1.00       141
           2       0.96      1.00      0.98       104

    accuracy                           0.99       275
   macro avg       0.99      0.96      0.97       275
weighted avg       0.99      0.99      0.99       275
```

Model 3: Data Augmentation for Enhanced Robustness

Challenge

We aimed to enhance the model's robustness by making it more capable of handling variations in input data.

Solution

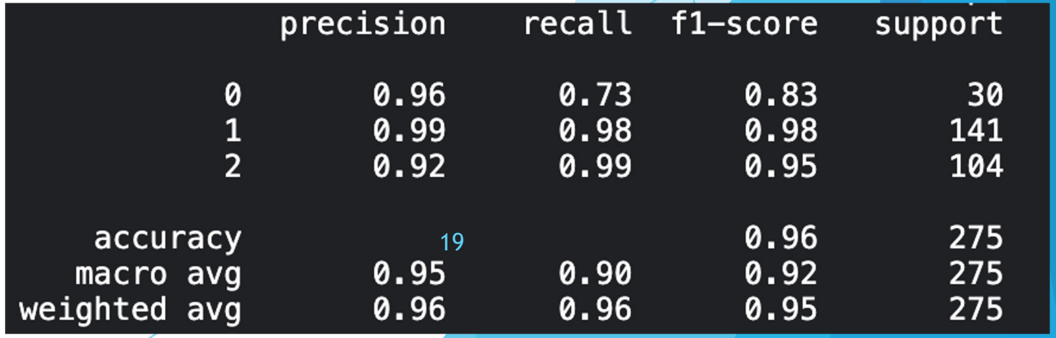
Data augmentation techniques were employed, including rotations, flips, and cropping, to artificially increase the diversity of our dataset.

Outcome

The augmented dataset was used to train our CNN, making the model more resilient to variations in input images.

Performance

The model achieved an accuracy of 93% on the validation data.



```
              precision    recall  f1-score   support

           0       0.96      0.73      0.83        30
           1       0.99      0.98      0.98       141
           2       0.92      0.99      0.95       104

    accuracy                           0.96       275
   macro avg       0.95      0.90      0.92       275
weighted avg       0.96      0.96      0.95       275
```

Conclusion

Based on the evaluation metrics obtained for each model on the test dataset, it appears that all three models (model1, model2, and model3) perform exceptionally well. Here's a summary of their performance:

Model 1:

Accuracy: 0.94

Precision (macro avg): 0.89

Recall (macro avg): 0.93

F1-score (macro avg): 0.91

Conclusion

Model 2:

Accuracy: 0.99

Precision (macro avg): 0.97

Recall (macro avg): 0.99

F1-score (macro avg): 0.98

Model 3:

Accuracy: 0.94

Precision (macro avg): 0.89

Recall (macro avg): 0.93

F1-score (macro avg): 0.91

Conclusion

Different Models have different use cases we got the best results on model1 and model2

We tried with both Model 1 and Model 2 both are giving us good results.

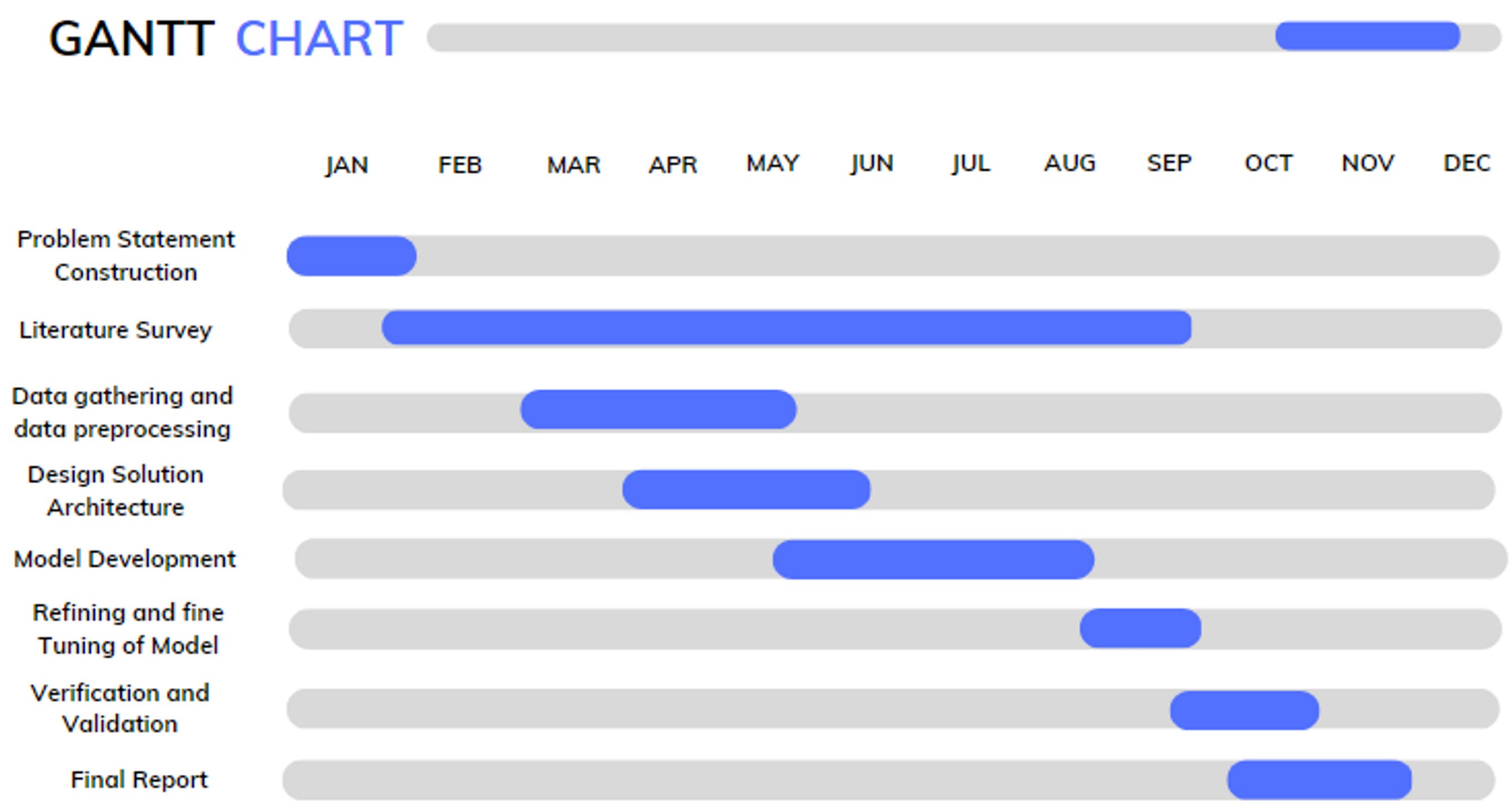
Expected Deliverables

Tabulate the individual contribution of the team members with the following,

a) Tasks/Modules assigned

Name	**Task**
Ayan Aggarwal	Image Preprocessing SMOTE Model Building
Akshat Bhandari	Data Collection and Exploration Image Preprocessing Model Building
Ayush Singh	Image Preprocessing Model Building Model Evaluation
Ankur Kumar Dubey	SMOTE Model Building Model Evaluation

Capstone (Phase-I & Phase-II) Project Timeline



References

References pertaining to your research

[1] - **Lung Cancer Prediction using Machine Learning: A Comprehensive Approach** - Syed Saba Raoof , M A. Jabbar, Syed Aley Fathima - 2020

[2] - **Lung cancer prediction and Stage classification in CT Scans Using Convolution Neural Networks -A Deep learning Model -** V.Deepa, P.Mohamed Fathimal - 2022

[3] - **Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier -** Janee Alam , Sabrina Alam , Alamgir Hossain -2022

[4] - **Isoline Based Image Colorization -**Adam Popowicz, Bogdan Smolk - 2014

References

Provide references pertaining to your research

[5] - **Lung Cancer Detection and Classification using CT Scan Image Processing** by Nusrat Nawreen , Umma Hany ,Tahmina Islam Department of Electrical and Electronic Engineering-2021

[6]-**Lung nodules: A comprehensive review on current approach and management** by Konstantinos Loverdos, Andreas Fotiadis, Chrysoula Kontogianni, Marianthi Iliopoulou, and Mina Gaga

[7]-**Lung Cancer Detection and Prediction of Cancer Stages Using Image Processing** byS.A.D.L.V. Senarathna ,S.P.Y.A.A. Piyumal ,R. Hirshan,W.G.C.W. Kumara-2021

[8]-**Image Acquisition and Pre-processing for Detection of Lung Cancer using Neural Network** by B C Kavitha; K B Naveen -2022

References

[9]-**Proposed methodology for Early Detection of Lung Cancer with low-dose CT Scan using Machine Learning** by Gagan Thakral, Sapna Gambhir , Nagender Aneja - 2022

[10]-**A Comparative Study of Image Segmentation Technique applied for Lung Cancer Detection**-Mohd Firdaus Abdullah; Muhammad Safwan Mansor; Siti Noraini Sulaiman; Muhammad Khusairi Osman-2019

Thank You