DON BOSCO INSTITUTE OF TECHNOLOGY

Premier Automobiles Road, Kurla (W), Mumbai-70

Approved by AICTE, Govt. of Maharashtra

&

Affiliated to the University of Mumbai

T.E. MINI PROJECT REPORT

CSM601 - Mini Project 2B

On

"EfficientNetB1 Model for Lung Cancer Detection using Biopsy Images"

Department of Computer Engineering

University of Mumbai

April 2024

CERTIFICATE

Project Title : **"Computer vision model for lung cancer detection using biopsy images"**

Organization : Don Bosco Institute of Technology

Address : Premier Automobiles
Road, Kurla (W),
Mumbai-70

Project Team Members:
1. Jess John
2. Figo Fernandez
3. Aaradhya Deotale

Internal Guide : **Prof. Dipti Jadhav**

_____ _____

INTERNAL GUIDE (s) **EXTERNAL GUIDE (s)**

_____

HEAD, COMPUTER ENGINEERING

ABSTRACT

Lung cancer is a pervasive and lethal disease, necessitating early and accurate diagnosis. The project focuses on leveraging Convolutional Neural Networks (CNNs) for automated lung cancer detection in histopathological biopsy images, addressing the challenges associated with timely and precise diagnosis. Manual assessment by pathologists is error-prone and time- consuming, emphasizing the need for an automated approach. The research encompasses a comprehensive scope, including improving interpretability, classifying specific lung cancer subtypes, real-time intraoperative analysis, and expansion to other cancer types.

The project's methodology involves using a pre-trained EfficientNet-based model for image classification. The model demonstrates its effectiveness in distinguishing between benign and malignant lung cancer cells, as evidenced by robust training results. Additionally, it offers the potential for personalization in the diagnosis and treatment of lung cancer.

The findings of this research indicate that machine learning models, particularly the EfficientNet-based architecture, can substantially enhance lung cancer detection accuracy. The model's ability to differentiate between cancer subtypes and the potential for real-time analysis during surgeries signifies a significant advancement in lung cancer diagnosis and treatment.

In conclusion, this project addresses the critical need for accurate and timely lung cancer diagnosis, offering a ray of hope in the fight against this devastating disease. The developed model has the potential to revolutionize the field of radiology and oncology, providing a valuable tool for medical professionals and contributing to improved patient outcomes.

TABLE OF CONTENTS

TABLE OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Network
DL	Deep Learning
AI	Artificial Intelligence
HOG	Histogram of Oriented Gradients (HOG)
GPU	Graphics Processing Unit

CHAPTER 1: INTRODUCTION

Lung cancer is a global health concern, and its early and accurate diagnosis is paramount in improving patient outcomes. The traditional manual assessment of lung cancer cells in biopsy images is time- consuming and prone to errors, creating a pressing need for a more efficient and precise solution. This project seeks to address this challenge by harnessing the power of Convolutional Neural Networks (CNNs) to develop an automated system for the detection and classification of lung cancer in histopathological biopsy images.

The scope of this project extends beyond mere classification; it aims to enhance interpretability for medical professionals, making the model's predictions more comprehensible. Moreover, the model is designed to classify specific lung cancer subtypes, allowing for more tailored treatment approaches. Real-time analysis capabilities are integrated to assist in intraoperative decision-making during surgeries, a significant advancement in the medical field. Collaboration with pathologists for continuous model improvement and expansion into the detection of other cancer types further highlights the project's significance.

With the high incidence of lung cancer, especially in India, the need for early detection is critical. The project not only addresses the challenge of diagnosis but also delves into understanding the unique epidemiological trends in the Indian population. By leveraging technology, this research endeavors to make a substantial contribution to the field of oncology, ultimately enhancing patient care and outcomes.

In conclusion, this project presents a holistic approach to lung cancer detection, integrating advanced technology with medical expertise. Its success will not only expedite the diagnostic process but also pave the way for more personalized treatments, offering hope in the battle against lung cancer and potentially influencing preventive measures to combat this deadly disease.

CHAPTER 2: LITERATURE SURVEY

2.1 RESEARCH PAPER 1:

The literature survey presents a detailed review of a paper titled "A Review on Diagnosis of Lung Cancer and Lung Nodules in Histopathological Images using Deep Convolutional Neural Network" authored by Shimna P K, Shirly Edward, and Roshini T V[10]. It emphasizes the critical importance of early detection in combating lung cancer, a significant global health concern. Traditional imaging techniques are critiqued, with the paper advocating for the adoption of deep convolutional neural networks (CNNs) due to their potential to enhance diagnostic accuracy. The summary outlines key aspects discussed in the paper, including the significance of early detection, limitations of conventional imaging methods, and the utility of CNNs in lung cancer diagnosis. Notably, it provides a comprehensive review of relevant literature on CNN applications in this domain, discusses pertinent datasets available for researchers, and explores the role of lung nodule segmentation in improving diagnosis accuracy. Furthermore, the paper encapsulates recent studies employing CNNs for lung cancer detection and proposes future research directions, highlighting the need for enhanced segmentation methods and differentiation of early-stage cancers. It concludes by emphasizing the pivotal role of CNNs in facilitating early detection and underscores the necessity for ongoing research endeavors. Additionally, exemplary points and shortcomings of the paper are outlined. Noteworthy commendations include the provision of a valuable dataset list for researchers and insightful recommendations for future work. However, the paper is critiqued for its lack of detailed technical methodologies and model descriptions in the reviewed works, minimal discussion on accuracy metrics, and oversight in critically analyzing the limitations of deep learning in lung cancer detection. Moreover, it misses providing insights into the real-world applicability of the discussed methodologies. Overall, the paper emerges as a relevant contribution to the field of medical image analysis, particularly in the context of lung cancer detection, offering a comprehensive overview of recent research, datasets, and methodologies pertaining to the utilization of deep learning, specifically CNNs, for this crucial purpose.

2.2 RESEARCH PAPER 2:

The literature survey provides insights into a study titled "Detection of Lung Cancer from Pathological Images Using CNN Model" authored by Siming Huang and Zexuan Zhang[7]. The study presents the development of a hybrid CNN-based model tailored for the detection of lung cancer, focusing on accurate classification through deep learning techniques. Notably, the model integrates various feature extraction methods, including inception_v3, HOG, and DAISY, to enhance its precision.

The summary highlights the utilization of the Lung and Colon Cancer Histopathological Image dataset for training and evaluation, achieving an impressive accuracy rate of 99.60% in classifying lung cancer. The study underscores the potential of hybrid deep learning approaches in enabling precise diagnosis, emphasizing the innovative fusion of traditional feature extraction with deep learning techniques.

Exemplary points commend the study for its innovative hybrid model, high accuracy, comprehensive evaluation metrics, and utilization of real-world datasets for clinical relevance. Additionally, the study visualizes results through confusion matrices and offers forward-looking discussions on potential future research directions.

However, shortcomings of the study include the lack of discussion on clinical validation and real-world deployment, limited exploration of the model's generalizability to other medical conditions, model complexity due to multiple feature extraction methods, and absence of ethical considerations in data usage. Furthermore, the study is critiqued for its limited comparison to existing state-of-the-art approaches and the absence of information regarding required c39omputational resources.

In terms of relevancy, the research holds significance in the domain of cancer detection, particularly in aiding early diagnosis through medical image analysis and deep learning techniques. The hybrid model developed in the study, which combines HOG and DAISY with inception_v3, contributes to advancing image analysis methods for lung cancer detection.

2.3 RESEARCH PAPER 3:

The literature survey delves into a paper titled "Deep Learning in Lung and Colon Cancer Classifications" authored by Krishna Mridha, MD, Iftekhar Islam, Shamin Ashfaq, and Dipayan Barua[8]. The paper introduces a Deep Learning (DL) framework designed for accurately classifying colon and lung cancer cells, achieving a high testing accuracy rate of 98.3%. It proposes that the developed model could significantly benefit medical experts by providing an accurate and automated system for diagnosing various forms of colon and lung cancers.

Exemplary points of the paper commend its extensive utilization of deep learning for medical image analysis, attainment of high sensitivity in detecting abnormalities, utilization of a diverse dataset sourced from multiple hospitals, integration of radiologist feedback to improve the AI model, and potential assistance to radiologists in early disease detection. Additionally, the study discusses implications for improving healthcare outcomes.

However, the paper is critiqued for its limited sample size of 12,500 images for training, high false positive rate particularly for colon scans, occasional misclassification of organ types (colon vs. lung), lack of external validation and generalizability, reliance on preprocessed data, and limited discussion on clinical relevance and real-world application.

In terms of relevancy, the paper contributes to the field of medical image classification, deep learning, cancer detection, data preprocessing, and accuracy analysis. It utilizes deep learning and image processing techniques to classify lung and colon cancer cells from histopathological images, aligning with projects involving similar methodologies and objectives.

2.4 RESEARCH PAPER 4:

The literature survey discusses a paper titled "Automatic Detection of Various Types of Lung Cancer Based on Histopathological Images Using a Lightweight End-to-End CNN Approach" authored by Ahmed S. Sakr[4]. The paper introduces a Deep Learning (DL) framework aimed at classifying different types of lung cancer cells based on histopathological images, achieving a high testing accuracy rate of over 99.5%. It posits that the proposed model holds the potential to assist medical experts by offering an accurate and automated system for diagnosing various forms of lung cancer.

Exemplary points of the paper highlight the presentation of results, including training, accuracy, and loss curves, along with the confusion matrix. The model's performance showcases an impressive accuracy exceeding 99.5%.

However, the paper is critiqued for its limitations, including the utilization of a small dataset that limits diversity representation, lack of external dataset validation for model performance, inadequate detailing of technical architecture and hyperparameters, limited comprehensive analysis of evaluation metrics, a brief discussion section without thorough result analysis, and insufficient information regarding code availability on GitHub.

In terms of relevancy, the paper's focus on lung cancer histopathological image classification using deep learning aligns with projects involving medical image analysis. It offers insights and methodologies that could be pertinent to similar endeavors aimed at enhancing the accuracy and efficiency of cancer diagnosis through automated systems leveraging deep learning techniques.

2.5 PROBLEM STATEMENT AND OBJECTIVE

The accurate differentiation between benign and malignant lung cancer cells within histopathological biopsy images is a formidable obstacle in the realm of medical diagnosis. The complexity of this task hinders the timely identification of specific lung cancer types, which, in turn, profoundly impacts patientcare and treatment outcomes.

The current reliance on manual assessment by pathologists not only consumes a significant amount of time but also introduces an inherent susceptibility to human error. This human-centric approach adds an element of subjectivity to the diagnosis, potentially leading to inconsistencies and variations in treatment recommendations. Urgency to address this issue cannot be overstated, as early and accurate diagnosis of lung cancer is paramount in improving patient survival rates and the effectiveness of treatment strategies. The development of a computer vision model for automated lung cancer detection using biopsy images is, therefore, a critical endeavor. By harnessing the power of artificial intelligence and deep learning techniques, this project seeks to enhance diagnostic accuracy, expedite the evaluation process, and, most importantly, elevate the overall quality of patient care outcomes.

Objectives of the Mini project are as follows: -

- Create a highly accurate machine learning model capable of distinguishing between benign and malignant lung cancer cells in histopathological images.
- Enable healthcare professionals to leverage the model's predictions for timely diagnosis, leading to tailored treatment plans and enhanced patient outcomes.
- Establish a versatile framework for further expansion, encompassing the potential to refine classification accuracy, incorporate subtype classification, and contribute to ongoing research in lung cancer detection

2.6 SCOPE OF THE MINI PROJECT

- Enhancing the model's interpretability to aid medical professionals in understanding the reasoning behind predictions.
- Extending the model to classify specific lung cancer subtypes, allowing for even more targeted treatment approaches.
- Integrating real-time analysis capabilities to assist in intraoperative decision-making during surgeries.
- Collaborating with pathologists to continuously improve the model's performance through iterative updates and training on diverse datasets.
- Expanding the framework to encompass other cancer types, thereby contributing to a comprehensive diagnostic tool for various malignancies.

CHAPTER 3: PROPOSED SYSTEM

3.1 ANALYSIS/FRAMEWORK/ALGORITHM

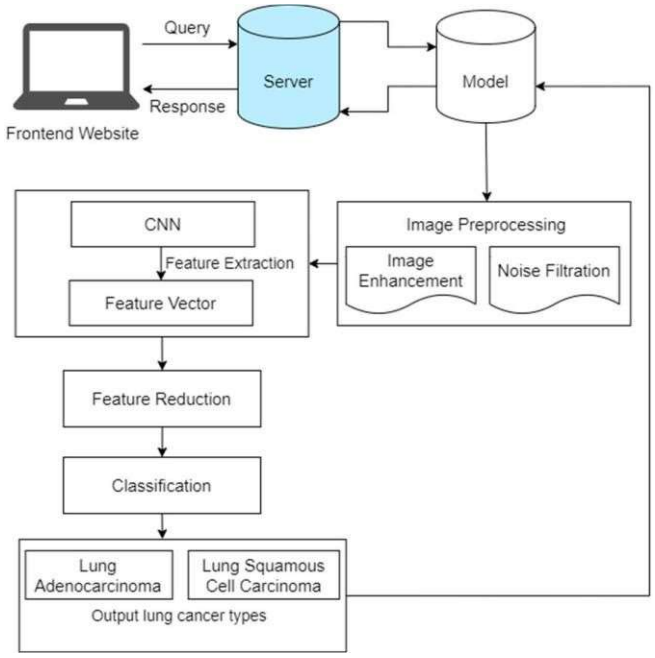

Fig 3.1.1 System Architecture

In our proposed system, we leverage state-of-the-art deep learning techniques, specifically Convolutional Neural Networks (CNNs), to address the critical challenge of accuratelydifferentiating between benign and malignant lung cancer cells in histopathological biopsy images. CNNs have shown remarkable success in image analysis, making them an ideal choice for this medical diagnosis task. We use a pre-trained EfficientNetB1 model, fine-tuned to our specific problem, to achieve high accuracy in classifying lung cancer cell images into three categories: "No Cancer," "Adenocarcinoma," and "Squamous Cell Carcinoma."

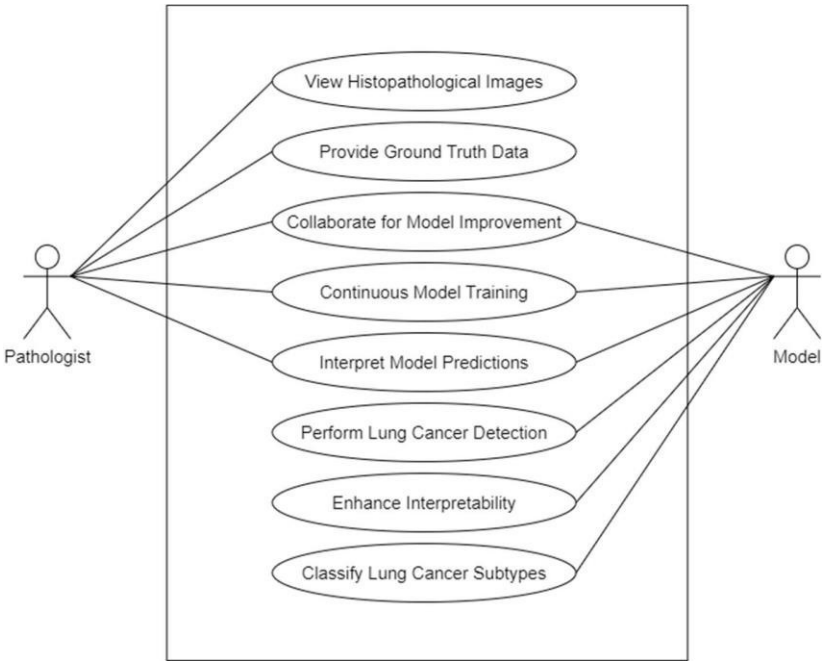

Fig 3.1.2 Use case diagram

3.2 DETAILS OF HARDWARE AND SOFTWARE

Our system's hardware requirements include a computer with sufficient processing power, ideally equipped with a modern GPU to accelerate deep learning computations. For software, we rely on Python, TensorFlow, and various deep learning libraries. We utilize TensorFlow for model development, training, and deployment. Additionally, we employ data preprocessing techniques and visualization tools for a comprehensive analysis of the lung cancer cell images.

3.3 DESIGN DETAILS

The design of our system encompasses a data pipeline for image loading, preprocessing, and augmentation. We split the dataset into training and validation sets to train and evaluate our model's performance. We fine-tune a pre-trained CNN model, EfficientNetB1, to adapt it to our specific classification task. The model consists of multiple layers, including convolutional and fully connected layers, to extract and learn features from the input images. We employ techniques such as batch normalization and dropout to improve model robustness.

3.4 METHODOLOGY

Our approach involves the following key steps:

1. **Data Collection:** Data Collection and Preprocessing will be done on a dataset of lung biopsy images, including both benign and malignant cases.
2. **Data Preprocessing:** We preprocess the images by resizing them to a uniform size (256x256 pixels) and normalizing pixel values. We also augment the dataset with transformations to enhance model generalization.
3. **Model Selection:** We select the EfficientNetB1 pre-trained model as our base architecture due to its efficiency and effectiveness in image classification tasks.
4. **Transfer Learning:** We fine-tune the pre-trained model on our dataset, retraining the model's top layers while keeping the initial layers frozen to capture general image features.
5. **Model Training:** We train the model on the training dataset, monitoring its performance on the validation set. We utilize the "Adam" optimizer and sparse categorical cross-entropy loss.
6. **Performance Evaluation:** We assess the model's performance using metrics like accuracy, loss, and confusion matrices. Our aim is to achieve high accuracy and minimize misclassifications.
7. **Visualization:** We use data visualization tools to inspect the model's predictions and gain insights into its decision-making process.

CHAPTER 4: IMPLEMENTATION PLAN AND IMPLEMENTATION

4.1 IMPLEMENTATION DISCUSSION-MODULE WISE

Our findings present a comprehensive approach to developing and evaluating a deep learning model for the classification of lung images into three distinct categories: "No Cancer," "Adenocarcinoma," and "Squamous Cell Carcinoma." The methodology encompasses several key stages, each aimed at ensuring robust model performance and providing insights into the classification process. Initially, the paper details the data preprocessing and exploration phase, wherein the lung image datasets are loaded and processed using TensorFlow's image dataset utilities.

The datasets are split into training and validation sets, and an analysis of class distribution is conducted to identify potential data imbalances. Visualizations, including bar plots, are employed to illustrate the distribution of images across different classes, facilitating a deeper understanding of the dataset's characteristics. Subsequently, the model construction and training strategy are described, focusing on the utilization of transfer learning with the EfficientNetB1 architecture as the base model. Additional dense layers are appended to the base model to facilitate classification. The training process is managed through the implementation of appropriate callbacks, including early stopping and learning rate reduction, to prevent overfitting and optimize model performance. Performance metrics such as accuracy and loss are monitored and visualized over 25 epochs to track the model's learning progress. Figure 2 illustrates the suggested model's architecture. Following model training, the paper discusses the evaluation phase, wherein the trained model is assessed using the validation dataset. Sample predictions are generated and visualized to provide qualitative insights into the model's performance. Additionally, quantitative analyses, including the generation of a confusion matrix and a classification report, offer a more in-depth understanding of the model's classification accuracy and potential areas for improvement. Moving on, we further explore avenues for further analysis and visualization, including the visualization of learning rate schedules and the generation of sample predictions on the validation dataset. These analyses contribute to a comprehensive understanding of the model's learning process and predictive capabilities, enhancing the interpretability and reliability of the classification results.

The data were partitioned using an 80-20 split for training and testing, respectively, employing Python 3.7. Key performance metrics, including accuracy, precision, recall, and F1-score, were prioritized for evaluating the system's generalization and classification capabilities. The dataset consisted of a total of 5000 images across three classes: No Cancer, Adenocarcinoma Class, and Squamous Cell Carcinoma Class.

The specific definitions of these metrics are elaborated below:

- Precision = $\frac{Tp}{Tp+FP}$
- Recall = $\frac{Tp}{Tp+Fn}$
- Accuracy = $\frac{Tp+Fn}{Tp+Tn+Fp+Fn}$
- F1-Score = $\frac{Tp}{Tp+FTp+0.5*(Fp+Fn)\,n}$

In the above equations, Tp stands for True Positive, Fp stands for False Positive, while Tn stands for True Negative and Fn stands for False Negative.

As for the EfficientNetB1 CNN architecture itself; it consists of three layers, two of them dense layers and one of them a functional layer. These layers are: - EfficientNetB1, Dense and Dense_1 layers. Details about the same can be seen in the figure below: -

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 efficientnetb1 (Functional)  (None, 1280)             6575239

 dense (Dense)               (None, 128)               163968

 dense_1 (Dense)             (None, 3)                 387

=================================================================
Total params: 6,739,594
Trainable params: 164,355
Non-trainable params: 6,575,239
_________________________________________________________________
```

Fig: 4.1.1: Architecture of EfficientNetB1 model

A thorough literature review was conducted to understand the current landscape of lung cancer detection using deep learning. Several research papers were considered that emphasize the importance of early detection and the role of Convolutional Neural Networks (CNNs) in this domain. This helps in acknowledging the limitations and strengths of these studies, and based off that minute analysis, insights for the implementation for this mini-project were considered.

As part of our project, we are extensively studying the specific tools, techniques, and methodologies used in these papers. We are also exploring relevant datasets for training and testing our model. This deeper understanding is guiding our implementation.

In our proposed work, we are designing a deep learning model that accurately classifies histopathological lung images into different categories, including "No Cancer," "Adenocarcinoma," and "Squamous Cell Carcinoma." Our model is based on a Convolutional Neural Network architecture.

We utilized mathematical modeling to define the architecture of our CNN. This included specifying the number of layers, filters, and neurons in the network. The choice of activation functions and loss functions was also crucial. These mathematical components shaped the behavior of our model.

Our input was histopathological lung images of a fixed size (e.g., 256x256 pixels). The output was the predicted class, indicating the type of lung cancer or a "No Cancer" prediction. This input-output mapping guided the training and testing phases of our model.

We used the Python programming language for implementing our deep learning model. Key tools and libraries that we relied on included TensorFlow, which is a popular framework for building and training neural networks. We leveraged pre-trained models like EfficientNetB1 as a base and fine-tuned them for our specific task.

Data Preparation:

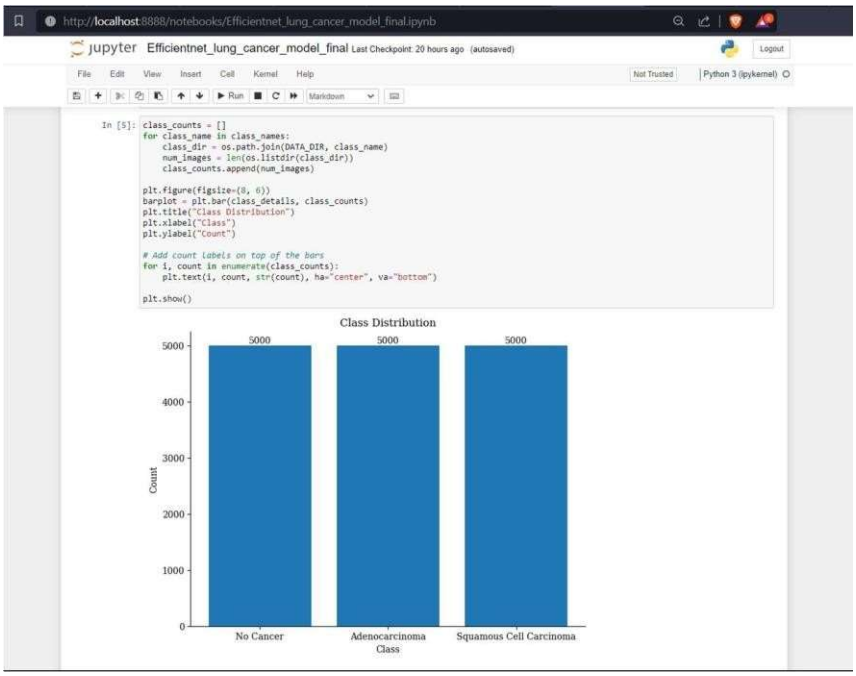

Fig 4.1.2 Data preprocessing done on the dataset

We gathered and preprocessed a dataset containing histopathological lung images. This dataset was split into training and validation sets to train and evaluate our model's performance.

Model Architecture:

Our model is based on a pre-trained EfficientNet architecture, which has proven effective in image classification tasks. We are adding a few additional layers, including fully connected layers, for fine-tuning the model.

Training and Evaluation:

We trained the model using the training dataset and monitored its performance using the validation set. We used metrics like accuracy, loss, and possibly others mentioned in the literature (precision, recall, F1-score) to assess the model's effectiveness.

Model Saving and Loading:

Once the model was trained and performed well, we saved its weights and architecture for future use. This allowed us to reuse the model without having to retrain it from scratch.

Visualization and Interpretation:

We visualized the model's predictions on sample images to ensure it was working correctly. We also visualized the confusion matrix to understand its performance in classifying different typesof lung cancer.

Deployment and Interaction:

After developing a reliable model, we are considering deploying it for real-world use. We are creating a user-friendly interface for interacting with the model, allowing medical professionals to upload and analyses histopathological images. The Gradio library is being used for creating such an interface.

4.2 IMPLEMENTATION (CODE AND SCREENSHOT OF IMPLEMENTATION)

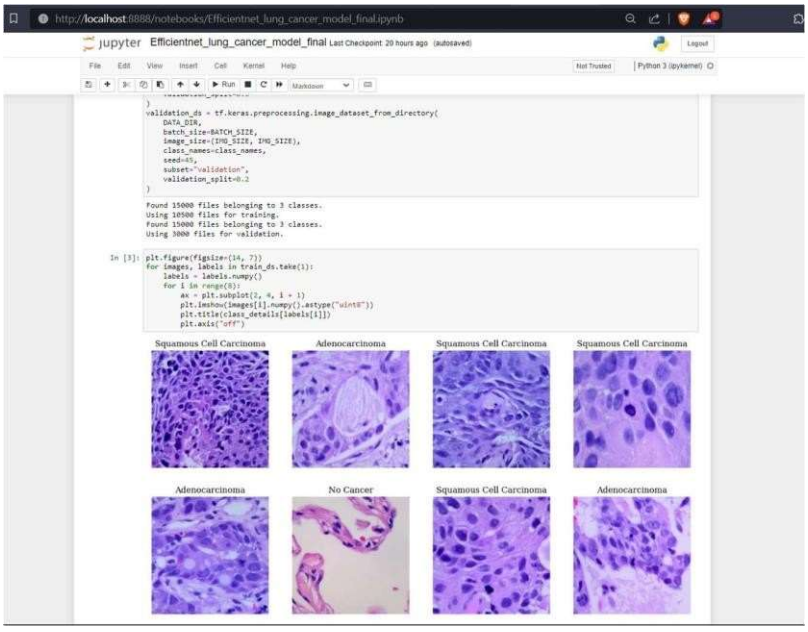

Fig 4.2.1 Dataset description

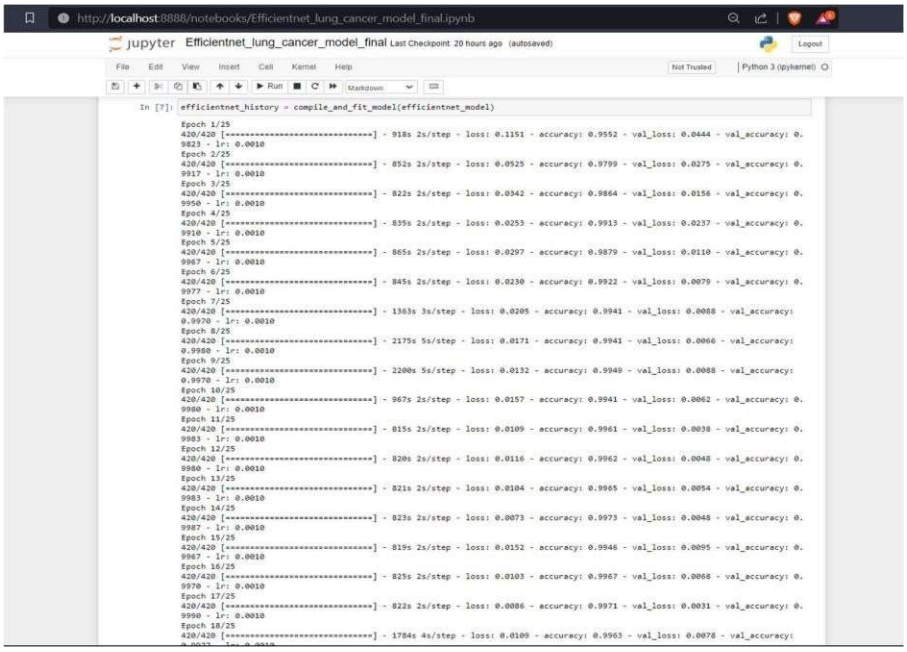

Fig 4.2.2 Model training

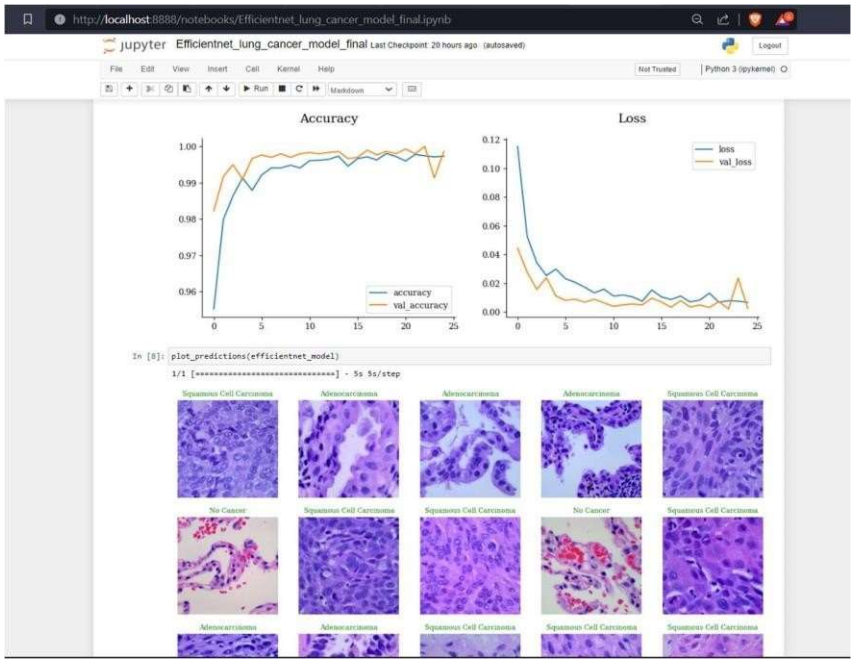

Fig 4.2.3 Model accuracy representation

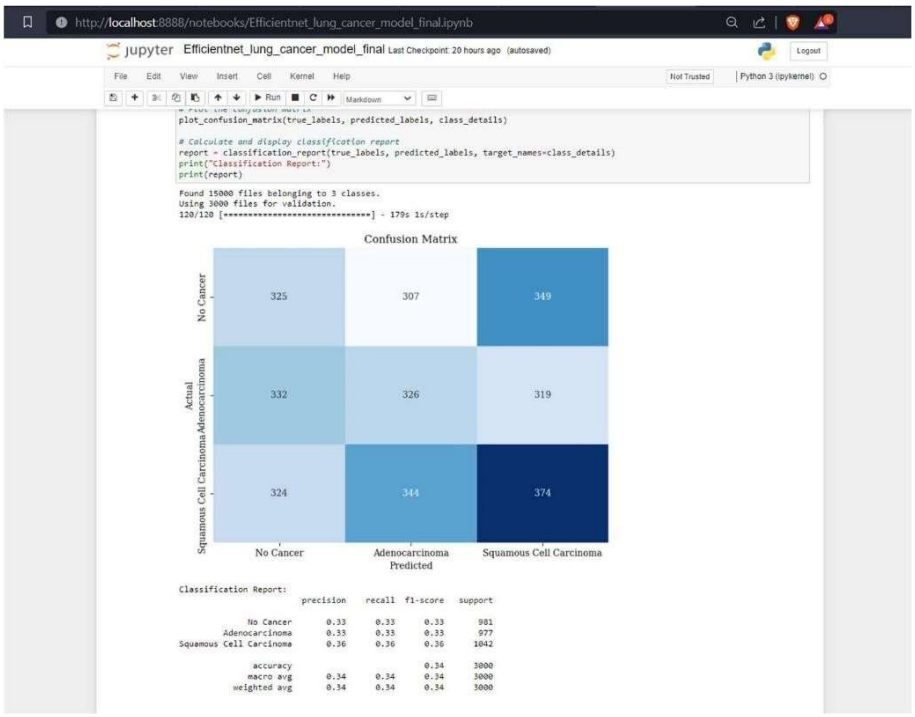

Fig 4.2.4 Confusion Matrix

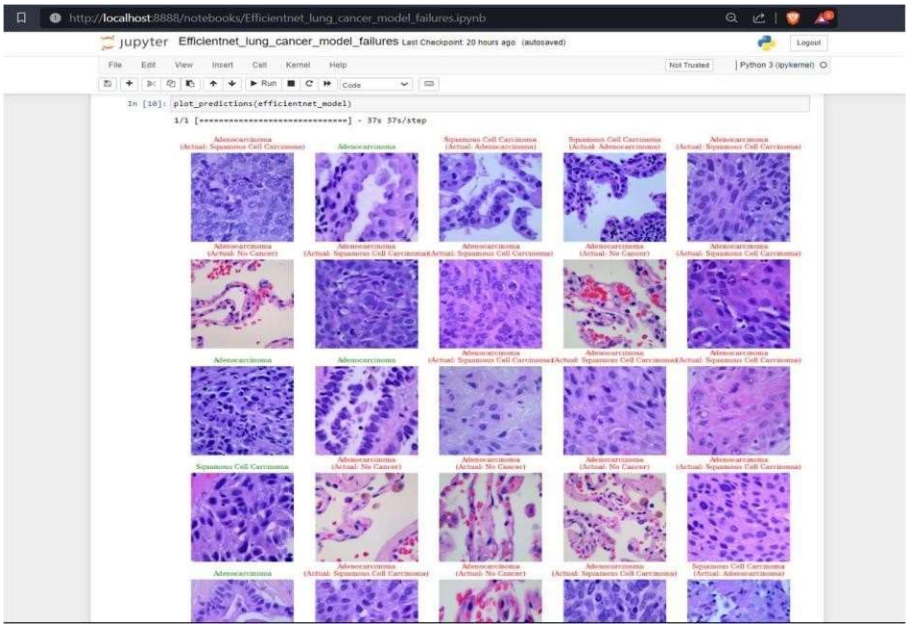

Fig 4.2.5 Inaccurately predicted exceptions

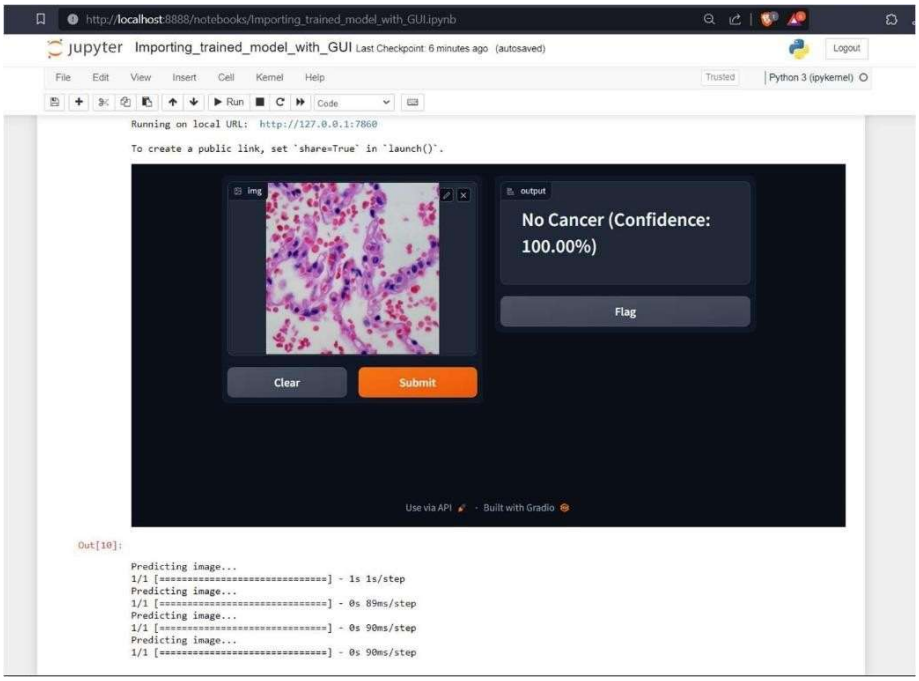

Fig 4.2.6 Test Case 1- Biopsy Image of cells having no Cancer

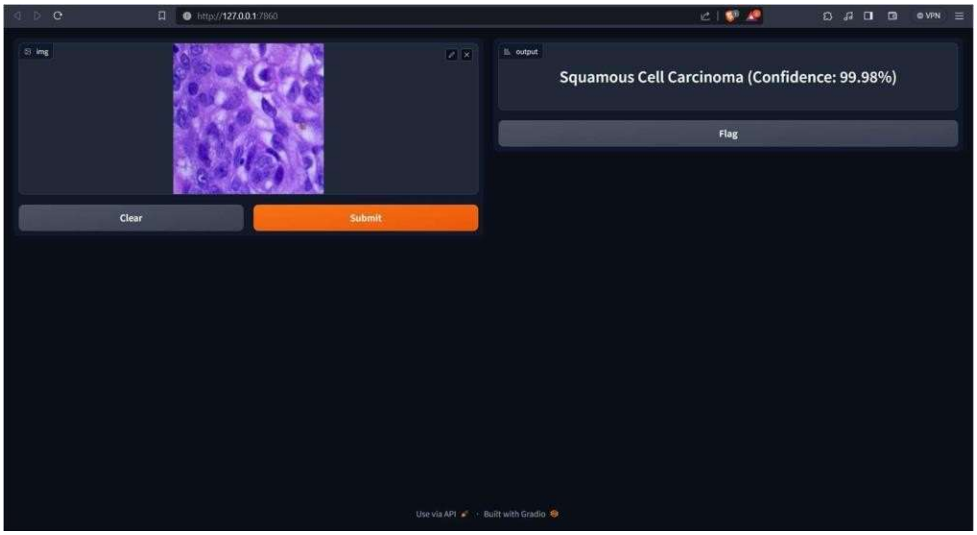

Fig 4.2.7 Test Case 2- Biopsy Image of cells having Squamous Cell Carcinoma

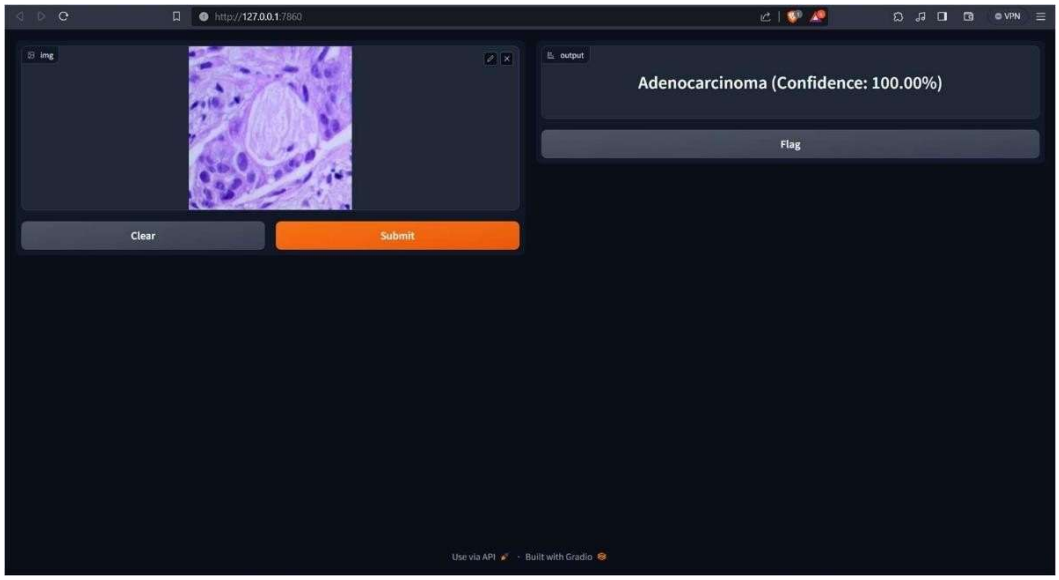

Fig 4.2.8 Test Case 3- Biopsy Image of cells having Adenocarcinoma

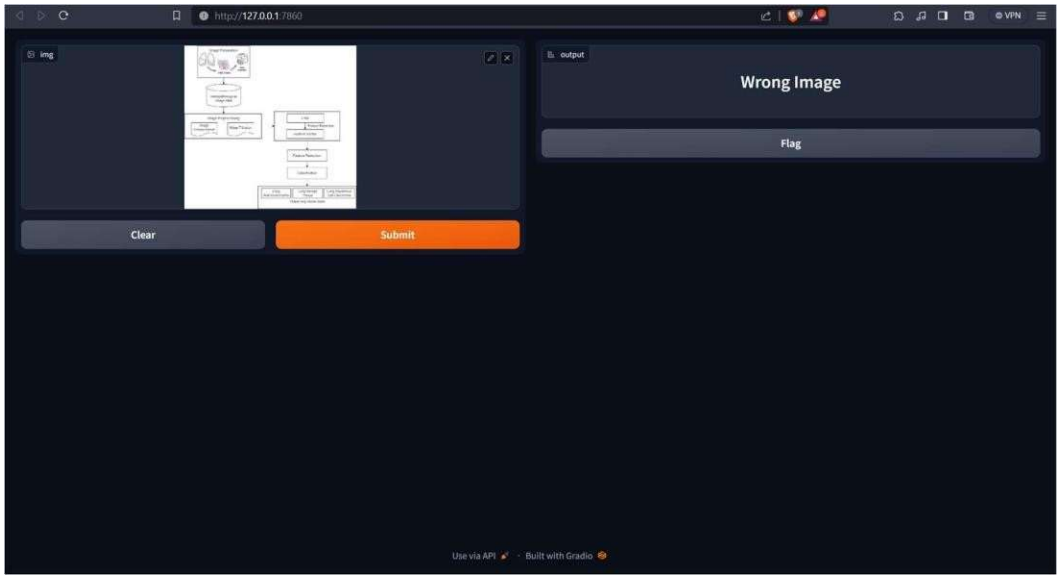

Fig 4.2.9 Test Case 4- Invalid input

CHAPTER 5: RESULTS AND DISCUSSION

In this section, we will be analyzing the results obtained in the implementation of three models viz. VGG16, ResNet50 and EfficientNetB1. These results will be compared with each other in terms of Accuracy, Loss and also their respective Confusion Matrices.

Our analysis reveals a distinct contrast in both accuracy and loss among the three models. The EfficientNetB1 model consistently demonstrates higher accuracy than the other two models, with ResNet50 closely following, although it experiences notable fluctuations in accuracy at the beginning and end of the training. Conversely, the VGG16 model exhibits an oscillating trend in accuracy, with an overall inconclusive pattern that gradually stabilizes towards the end of the training period.

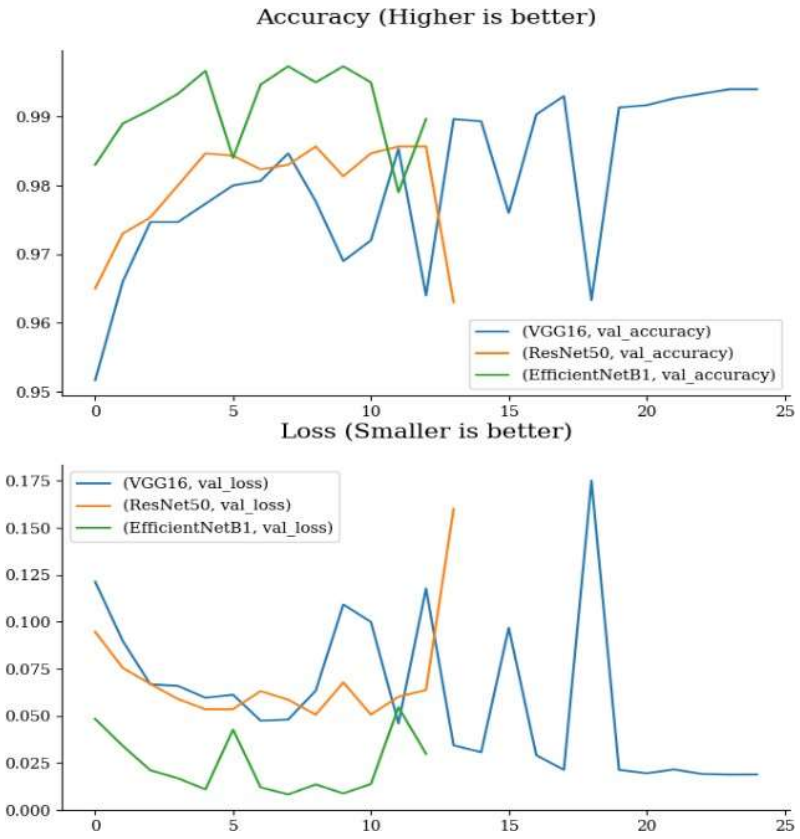

Fig 5.1 Comparison of Accuracy and Loss for all three models

Now, as for the comparison of the Confusion Matrices of all the three models, we see that barring some insignificant differences, all three models successfully align with the predicted and the true labels; indicating that they were largely successful in classifying the lung cancer biopsy images into the Adenocarcinoma, Squamous Cell Carcinoma and Benign Tissue labels.

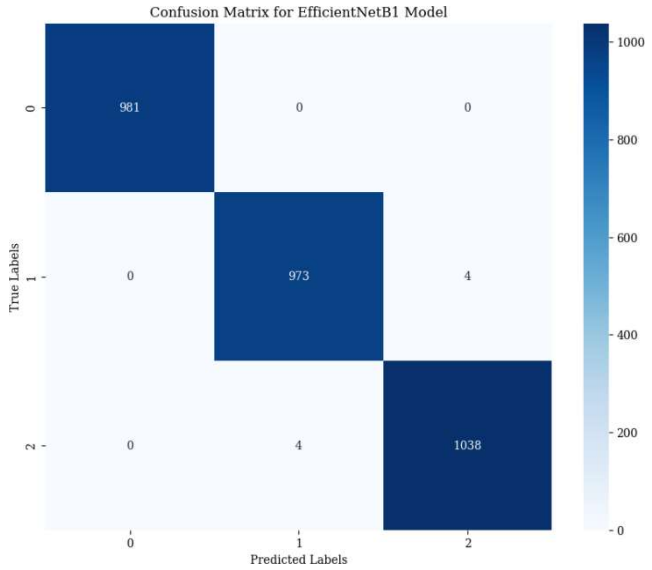

Fig 5.2 Confusion Matrix of EfficientNetB1 Model

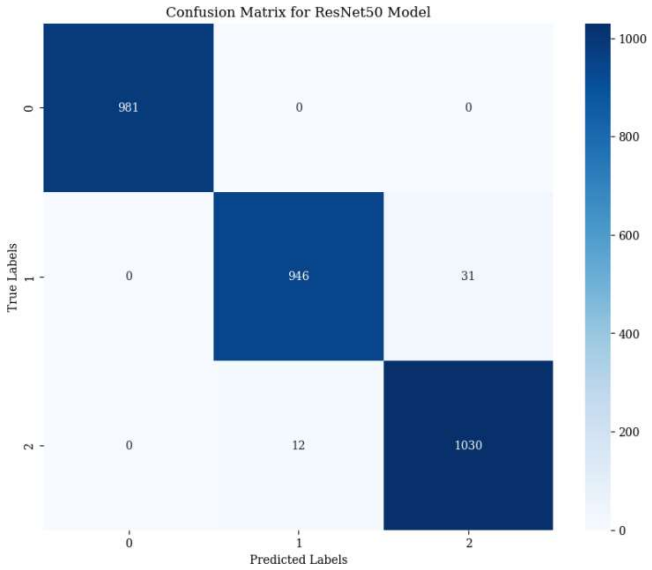

Fig 5.3 Confusion Matrix of ResNet50 Model

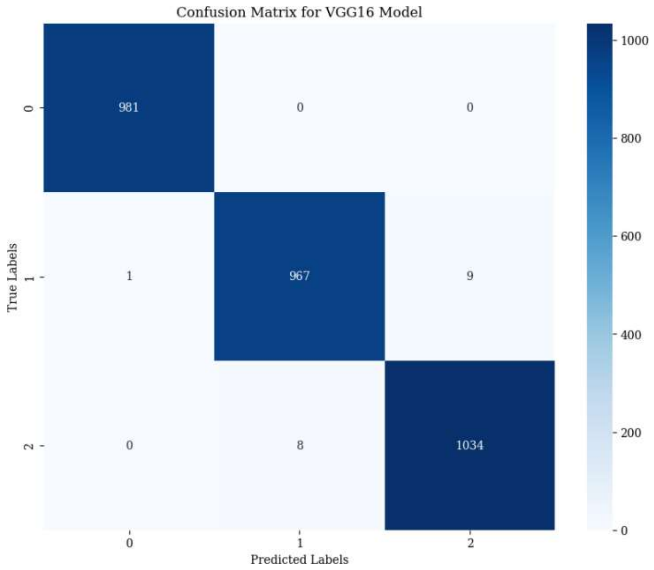

Fig 5.4 Confusion Matrix of VGG16 Model

CONCLUSION

In summary, this project addresses the critical challenge of accurately distinguishing between benign and malignant lung cancer cells in histopathological biopsy images. The proposed solution leverages Convolutional Neural Networks (CNNs) to automate the detection process, potentially expediting diagnoses and improving patient outcomes. The project also focuses on making the model's predictions interpretable, aiding medical professionals in understanding the reasoning behind its decisions.

To categorize lung tissue images from the Lung and Colon Cancer Histopathological Image dataset (LC25000) [12], a hybrid deep learning model is proposed in this study. The 768 x 768 photos that were obtained from LC25000 were subsequently scaled to 224 x 224 so that they could be used as model input. A feature extractor and a classifier make up the model.

The main contribution of this work is the suggestion of a lightweight deep-learning approach for end-to-end CNN-based lung cancer diagnosis using EfficientNetB1 model. After comparison with other models such as ResNet50 and VGG16, EfficientNetB1 was deemed to be more effective in all parameters viz. Accuracy, Loss, Precision, Recall, etc. The efficacy of the suggested system is evaluated and contrasted with other methods in this field using a database of histopathology images. According to the results, our method outperformed most previous deep-learning lung cancer diagnosis methods. Our model's highest accuracy is

0.995 percent. Compared to earlier deep models, the proposed method for diagnosing lung cancer is more robust and efficient. In the future, we plan to investigate our deep model's performance on more datasets. In addition, we may apply optimization strategies in conjunction with our deep model to identify the most optimally recovered deep features. As lung cancer poses a significant health concern worldwide, especially in India, where the incidence is rising, early detection is paramount. By combining technology with medical expertise, this project strives to expedite diagnoses, improve patient outcomes, and influence preventive measures, offering hope in the fight against this deadly disease.

REFERENCES

JOURNAL PAPERS

[1] D. P. P, J. K. Radhakrishnan, K. S. Aravind, P. R. Nambiar and N. Sampath, "Detection of Non-small cell Lung Cancer using Histopathological Images by the approach of Deep Learning," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-11, doi: 10.1109/CONIT55038.2022.9847945. U. Maheshwari, B. V. Kiranmayee and C. Suresh, "Diagnose Colon and Lung Cancer Histopathological Images Using Pre-Trained Machine Learning Model,"

[2] U. Maheshwari, B. V. Kiranmayee and C. Suresh, "Diagnose Colon and Lung Cancer Histopathological Images Using Pre-Trained Machine Learning Model," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 1078-1082, doi: 10.1109/IC3I56241.2022.10073184. A. S. Sakr, "Automatic Detection of Various Types of Lung Cancer Based on Histopathological Images Using a Lightweight End-to-End CNN Approach," 2022 20th International Conference on Language Engineering (ESOLEC), Cairo, Egypt, 2022, pp. 141-146, doi: 10.1109/ESOLEC54569.2022.10009108. J. Williams, "Narrow-Band Analyzer," PhD dissertation, Dept. of Electrical Eng., Harvard Univ., Cambridge, Mass., 1993. (Thesis or dissertation)

[3] R. D. Mohalder, J. P. Sarkar, K. A. Hossain, L. Paul and M. Raihan, "A Deep Learning Based Approach to Predict Lung Cancer from Histopathological Images," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), Khulna, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641341. L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, no.4, pp. 193-218, Apr. 1985. (Journal or magazine citation)

[4] A. S. Sakr, "Automatic Detection of Various Types of Lung Cancer Based on Histopathological Images Using a Lightweight End-to-End CNN Approach," 2022 20th International Conference on Language Engineering (ESOLEC), Cairo, Egypt, 2022, pp. 141-146, doi: 10.1109/ESOLEC54569.2022.10009108. J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746. (Preprint)

[5] J. Cañada, E. Cuello, L. Téllez, J. M. García, F. J. Velasco and J. Cabrera, "Assistance to lung cancer detection on histological images using Convolutional Neural Networks," 2022 E-Health and Bioengineering Conference (EHB), Iasi, Romania, 2022, pp. 1-4, doi: 10.1109/EHB55594.2022.9991400.

[6] D. P. P, J. K. Radhakrishnan, K. S. Aravind, P. R. Nambiar and N. Sampath, "Detection of Non-small cell Lung Cancer using Histopathological Images by the approach of Deep Learning," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-11, doi: 10.1109/CONIT55038.2022.9847945.

[7] M. Chen, S. Huang, Z. Huang and Z. Zhang, "Detection of Lung Cancer from Pathological Images Using CNN Model," 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 2021, pp. 352-358, doi: 10.1109/CEI52496.2021.9574590.

[8] K. Mridha, M. I. Islam, S. Ashfaq, M. A. Priyok and D. Barua, "Deep Learning in Lung and Colon Cancer classifications," 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), Dehradun, India, 2022, pp. 1-6, doi: 10.1109/ICACCM56405.2022.10009311.

[9] D. Nannapaneni, V. R. S. V. Saikam, R. Siddu, V. M. Challapalli and V. Rachapudi, "Enhanced Image-based Histopathology Lung Cancer Detection," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 620-625, doi: 10.1109/ICCMC56507.2023.10084247.

[10]P. K. Shimna, A. Shirly Edward and T. V. Roshini, "A Review on Diagnosis of Lung Cancer and Lung Nodules in Histopathological Images using Deep Convolutional Neural Network," 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India, 2023, pp. 1-4, doi: 10.1109/ICAIA57370.2023.10169738.

WEBSITES:

https://journals.lww.com/ijmr/fulltext/2022/02000/a_clinicoepidemiological_profile_of_lung_cancers.7.aspx (Accessed on 3/10/2023)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991145/(Accessed on 3/10/2023)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991146/(Accessed on 7/10/2023)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7220483/(Accessed on 11/10/2023)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6982605/(Accessed on 30/9/2023)

https://www.livemint.com/news/india/nithari-killings-investigation-was-botched-up-what-allahabad-high-court-said-top-quotes-11697469981237.html (Accessed on 7/10/2023)

DATASET:

https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images (Accessed on 13/8/2023)

APPENDIX

APPENDIX I - MODEL CODE:

(Note: The following code is in Python language, and is implemented in an. ipynb format)

```
import os


import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import tensorflow as tf
from tensorflow.keras import layers



SEED = 15243
np.random.seed(SEED)
os.environ["PYTHONHASHSEED"] = str(SEED)
tf.random.set_seed(SEED)


plt.rc("axes.spines", right=False, top=False)
plt.rc("font", family="serif")
BATCH_SIZE = 25
DATA_DIR = ("archive/"
"lung_colon_image_set/lung_image_sets")
IMG_SIZE = 256
MAX_EPOCHS = 25
```

```
class_names = ["lung_n", "lung_aca", "lung_scc"]
class_details = ["No Cancer", "Adenocarcinoma", "Squamous Cell Carcinoma"]

train_ds = tf.keras.preprocessing.image_dataset_from_directory(
DATA_DIR,
batch_size=BATCH_SIZE,
image_size=(IMG_SIZE, IMG_SIZE),
class_names=class_names,
seed=45,
subset="training",
validation_split=0.3
)
validation_ds = tf.keras.preprocessing.image_dataset_from_directory(
DATA_DIR,
batch_size=BATCH_SIZE,
image_size=(IMG_SIZE, IMG_SIZE),
class_names=class_names,
seed=45,
subset="validation",
validation_split=0.2
)
plt.figure(figsize=(14, 7))
for images, labels in train_ds.take(1):
labels = labels.numpy()
```

```
for i in range(8):
ax = plt.subplot(2, 4, i + 1)
plt.imshow(images[i].numpy().astype("uint8"))
plt.title(class_details[labels[i]])
plt.axis("off")
# Cache and prefetch data for faster training
AUTOTUNE = tf.data.AUTOTUNE
train_ds = train_ds.cache(".cached-data").prefetch(buffer_size=AUTOTUNE)
validation_ds = validation_ds.cache().prefetch(buffer_size=AUTOTUNE)


def compile_and_fit_model(model: tf.keras.Sequential) -> tf.keras.callbacks.History:
model.compile(
optimizer="adam",
loss="sparse_categorical_crossentropy",
metrics=["accuracy"]
)
reduce_lr = tf.keras.callbacks.ReduceLROnPlateau(
monitor="val_loss",
factor=0.2,
patience=5,
min_lr=0.001
)
history = model.fit(
```

```
63. train_ds,
64. validation_data=validation_ds,
65. epochs=MAX_EPOCHS,
66. callbacks=[reduce_lr]
67.)
68. performance_df = pd.DataFrame(history.history)
69. fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(12, 4))
70. for ax, metric in zip(axes.flat, ["accuracy", "loss"]):
71. performance_df.filter(like=metric).plot(ax=ax)
72. ax.set_title(metric.title(), size=15, pad=20)
73. return history


74. def plot_predictions(model: tf.keras.Sequential) -> None:
75. plt.figure(figsize=(14, 14))
76. for images, labels in train_ds.take(1):
77. labels = labels.numpy()
78. predicted_labels = np.argmax(model.predict(images), axis=1)
79. for i, (actual, pred) in enumerate(zip(predicted_labels, labels)):
80. ax = plt.subplot(5, 5, i + 1)
81. plt.imshow(images[i].numpy().astype("uint8"))
82. if actual == pred:
83. plt.title(class_details[labels[i]], color="green", size=9)
84. else:
```

```
plt.title(f"{class_details[predicted_labels[i]]}\n"
+ f"(Actual: {class_details[labels[i]]})",
color="red", size=9)
plt.axis("off")
class_counts = []
for class_name in class_names:
class_dir = os.path.join(DATA_DIR, class_name)
num_images = len(os.listdir(class_dir))
class_counts.append(num_images)

plt.figure(figsize=(8, 6))
barplot = plt.bar(class_details, class_counts)
plt.title("Class Distribution")
plt.xlabel("Class")
plt.ylabel("Count")

# Add count labels on top of the bars
    for i, count in enumerate(class_counts):
    plt.text(i, count, str(count), ha="center", va="bottom")

    plt.show()
    prerained_efficientnet_base = tf.keras.applications.efficientnet.EfficientNetB1(
    include_top=False, weights="imagenet", pooling="avg",
    )
    prerained_efficientnet_base.trainable = False
```

```
efficientnet_model = tf.keras.Sequential([
layers.Input(shape=(IMG_SIZE, IMG_SIZE, 3)),
prerained_efficientnet_base,
layers.Dense(128, activation="relu"),
layers.Dense(3, activation="softmax")
])
efficientnet_model.summary()
efficientnet_history = compile_and_fit_model(efficientnet_model)
plot_predictions(efficientnet_model)
from sklearn.metrics import confusion_matrix, classification_report
import seaborn as sns

# Function to plot a confusion matrix
def plot_confusion_matrix(y_true, y_pred, class_names):
cm = confusion_matrix(y_true, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False,
xticklabels=class_names, yticklabels=class_names)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```

```
# Load test data
test_ds = tf.keras.preprocessing.image_dataset_from_directory(
DATA_DIR,
batch_size=BATCH_SIZE,
image_size=(IMG_SIZE, IMG_SIZE),
class_names=class_names,
seed=45,
subset="validation",
validation_split=0.2
)

# Extract true labels from the test dataset
true_labels = []
for images, labels in test_ds:
true_labels.extend(labels.numpy())

# Predict labels using your trained model
predicted_labels = efficientnet_model.predict(test_ds)
predicted_labels = np.argmax(predicted_labels, axis=1)

# Plot the confusion matrix
plot_confusion_matrix(true_labels, predicted_labels, class_details)

# Calculate and display classification report
report = classification_report(true_labels, predicted_labels,
target_names=class_details)
```

```
print("Classification Report:")
print(report)
#save the model
efficientnet_model.save_weights("efficientnet_model_weights.h5")
loaded_model = tf.keras.Sequential([
tf.keras.layers.Input(shape=(IMG_SIZE, IMG_SIZE, 3)),
prerained_efficientnet_base, # Reuse the pretrained base
tf.keras.layers.Dense(128, activation="relu"),
tf.keras.layers.Dense(3, activation="softmax")
])
loaded_model.load_weights("efficientnet_model_weights.h5")
loaded_model.compile(optimizer="adam", loss="sparse_categorical_crossentropy",
metrics=["accuracy"])
# Recreate the model with the same architecture
loaded_model = tf.keras.Sequential([
tf.keras.layers.Input(shape=(IMG_SIZE, IMG_SIZE, 3)),
prerained_efficientnet_base, # Reuse the pretrained base
tf.keras.layers.Dense(128, activation="relu"),
tf.keras.layers.Dense(3, activation="softmax")
])

# Load the model weights from the first notebook
loaded_model.load_weights("efficientnet_model_weights.h5")

# Compile the loaded model
```

```
loaded_model.compile(optimizer="adam", loss="sparse_categorical_crossentropy", metrics=["accuracy"])

# Now, you can use the loaded model for inference and visualization
result = loaded_model.predict(validation_ds)

# Call your visualization function (e.g., plot_predictions) using the loaded_model
plot_predictions(loaded_model)
import tensorflow as tf
import gradio as gr
import numpy as np
import cv2

# Assuming 'loaded_model' is defined and loaded elsewhere in your code
# Define your 'predict_image' function as follows:

def predict_image(img):
print("Predicting image...")

# Resize the image to (256, 256) and convert to a 3D array
img_copy = img.copy()
img_3d = cv2.resize(img_copy, (256, 256))
img_3d = np.array(img_3d).reshape(-1, 256, 256, 3)

# Load your model here (replace 'loaded_model' with your loaded model)
prediction = loaded_model.predict(img_3d)[0]

class_labels = ["No Cancer", "Adenocarcinoma", "Squamous Cell Carcinoma"]
```

```
# Get the index of the class with the highest probability
predicted_class_index = np.argmax(prediction)

if predicted_class_index < len(class_labels):
predicted_class = class_labels[predicted_class_index]
confidence_percentage = prediction[predicted_class_index] * 100
if confidence_percentage >= 99.5:
return f"{predicted_class} (Confidence: {confidence_percentage:.2f}%)"
else:
return "Wrong Image"
else:
return "Wrong Image"

image = gr.inputs.Image(shape=(250, 250))
label = gr.outputs.Label()

gr.Interface(fn=predict_image, inputs=image, outputs=label, debug=True).launch()
```

APPENDIX II -ADDITIONAL IMPLEMENTATION DETAILS:

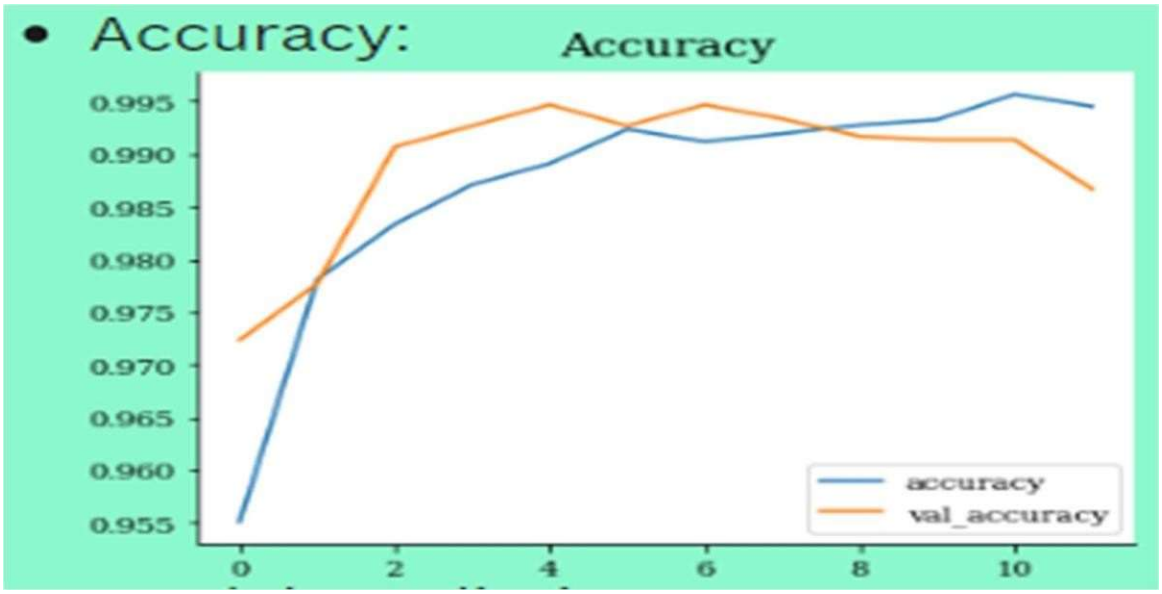

Fig A.II.1 Implementation Accuracy

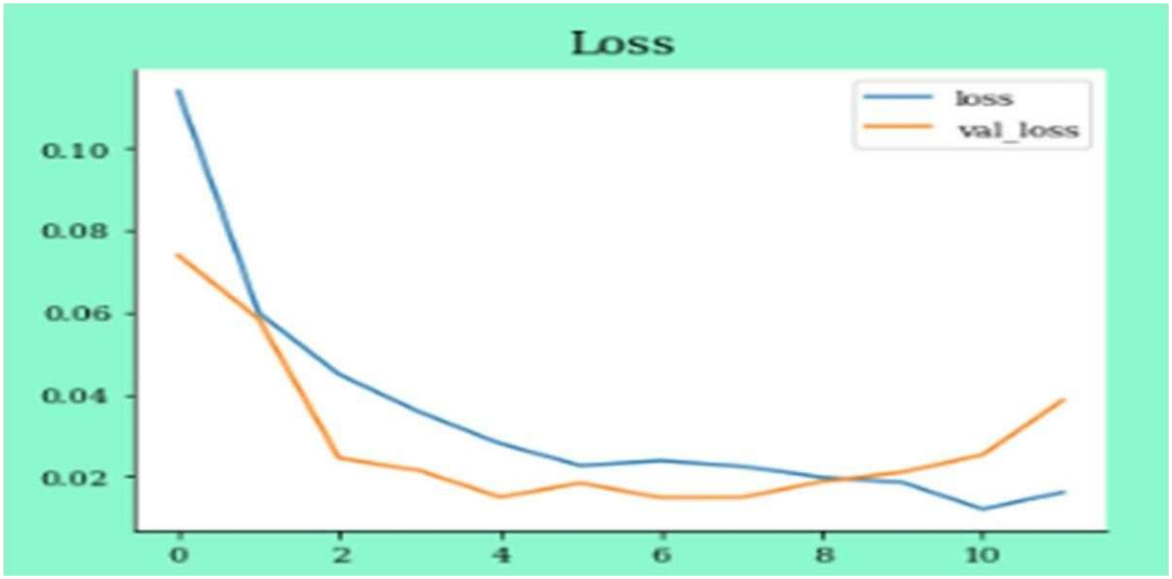

Fig A.II.2 Implementation Loss

APPENDIX III -LUNG CANCER

Lung cancer, a malignant neoplasm originating in the bronchi, bronchioles, or alveoli of thelungs, is a complex and heterogeneous disease with profound implications for both public health and clinical medicine. It is characterized by the uncontrolled proliferation of abnormal cells within lung tissues, leading to the formation of tumors. These tumors can obstruct the airways, compromise lung function, and metastasize to distant organs, significantly compromising patient prognosis.

Lung cancer is primarily categorized into two major histological types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), each with its unique characteristics and therapeutic approaches. NSCLC, accounting for approximately 85% of all cases, includes subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. SCLC, although less common, is known for its highly aggressive nature and rapid growth.

The etiology of lung cancer is multifaceted. The most well-established risk factor is tobacco smoking, which is responsible for the majority of lung cancer cases worldwide. Exposure to environmental carcinogens, such as radon, asbestos, and air pollutants, also contributes significantly to disease development, particularly in non-smokers. Furthermore, genetic factorsplay a role, with certain genetic mutations predisposing individuals to lung cancer. Clinical presentation varies but often includes symptoms such as persistent cough, hemoptysis, chest pain, shortness of breath, and weight loss. However, lung cancer frequently remains asymptomaticuntil advanced stages, making early diagnosis a challenge. Diagnostic procedures, such as chest X- rays, CT scans, bronchoscopy, and tissue biopsies, are employed to confirm the presence of lung cancer and determine its histological type and stage.

The management of lung cancer is multifaceted and dependent on the disease stage, histology, and patient's overall health. Treatment modalities encompass surgery, radiation therapy, chemotherapy, targeted therapy, and immunotherapy. Surgical resection with curative intent is often pursued for early-stage NSCLC, while advanced cases typically require a combination of chemotherapy, targeted therapy, and immunotherapy. The emergence of targeted therapies and immunotherapies has revolutionized the treatment landscape, offering new hope for patients with specific molecular alterations and improving survival rates.

In conclusion, lung cancer represents a critical public health challenge and a formidable clinical entity. Its multifactorial etiology, varied clinical presentation, and complex treatment strategies necessitate a multidisciplinary approach involving oncologists, surgeons, radiologists, and pathologists. Research into the molecular underpinnings of lung cancer continues to advance, promising new avenues for early detection and more effective, personalized treatment options.

Adenocarcinoma: Definition and Medical Analysis

Adenocarcinoma, a predominant histological subtype of non-small cell lung cancer (NSCLC), derives its name from the glandular cells in the lung where it typically originates. It is characterized by the uncontrolled growth of malignant cells in the alveoli and peripheral lung tissues, making it the most prevalent form of lung cancer, especially in non-smokers and young individuals. Adenocarcinoma is known for its diverse spectrum of genetic mutations and often presents as a solitary nodule or as multifocal, ground-glass opacities on imaging studies. This histological subtype is more likely to occur in the periphery of the lungs, and its slow growth can sometimes result in delayed diagnosis, contributing to an advanced disease stage at presentation.

Adenocarcinoma frequently carries specific genetic mutations, such as epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) rearrangements, which have profound therapeutic implications. Targeted therapies, including tyrosine kinase inhibitors (TKIs), have shown promising results in the management of adenocarcinoma with these mutations. In cases where surgical resection is feasible, curative intent surgery is often the preferred approach for early-stage disease. However, advanced adenocarcinoma may necessitate a combination of chemotherapy, targeted therapy, and immunotherapy.

Squamous Cell Carcinoma: Definition and Medical Analysis

Squamous cell carcinoma, another significant subtype of non-small cell lung cancer (NSCLC), originates from the squamous cells lining the bronchial tubes. It is characterized by the formation of keratinizing, stratified squamous epithelium, often leading to central tumor masses within the lung. Squamous cell carcinoma is typically associated with a history of tobacco smoking and tends to present with symptoms such as cough, hemoptysis, and post-obstructive pneumonia. Radiographically, it may manifest as a centrally located mass, often obstructing the bronchial tree and causing atelectasis.

The treatment approach for squamous cell carcinoma is influenced by the central location of the tumors. Surgical resection is often considered as a viable option in early-stage cases, while more advanced disease may require a combination of surgery, radiation therapy, and chemotherapy. Targeted therapies have not shown the same level of success in squamous cell carcinoma as in adenocarcinoma, mainly due to a lack of specific driver mutations. Immunotherapy, particularly immune checkpoint inhibitors, has emerged as a promising option in the treatment of squamous cell carcinoma, with some patients experiencing prolonged responses.

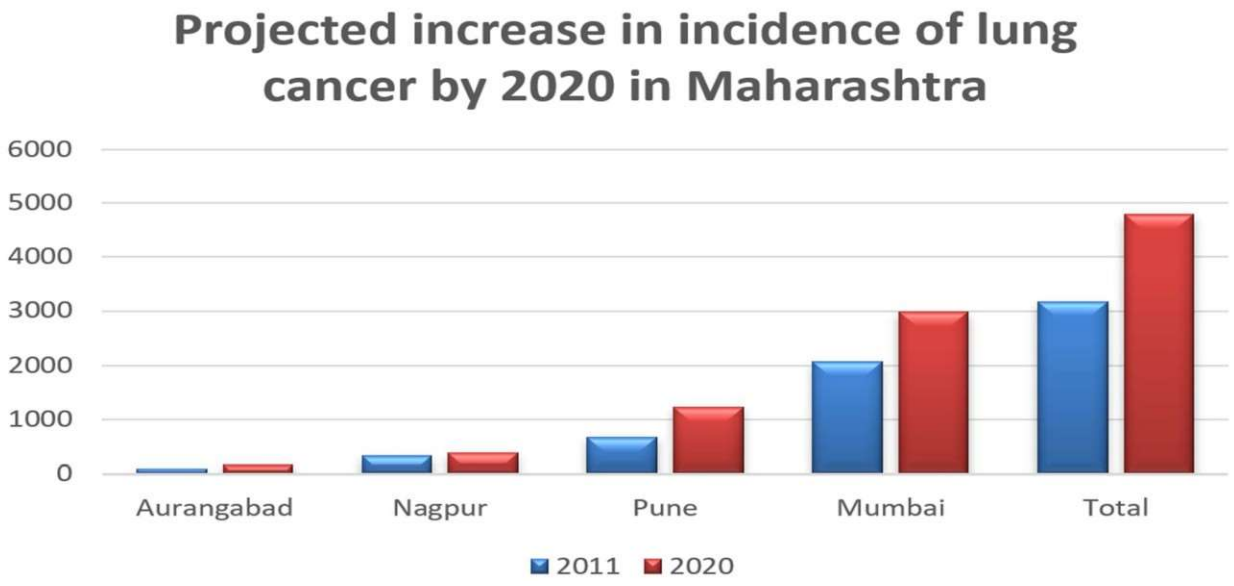

Fig A.III.1 Projected Increase in Lung Cancer, Maharashtra (2020)

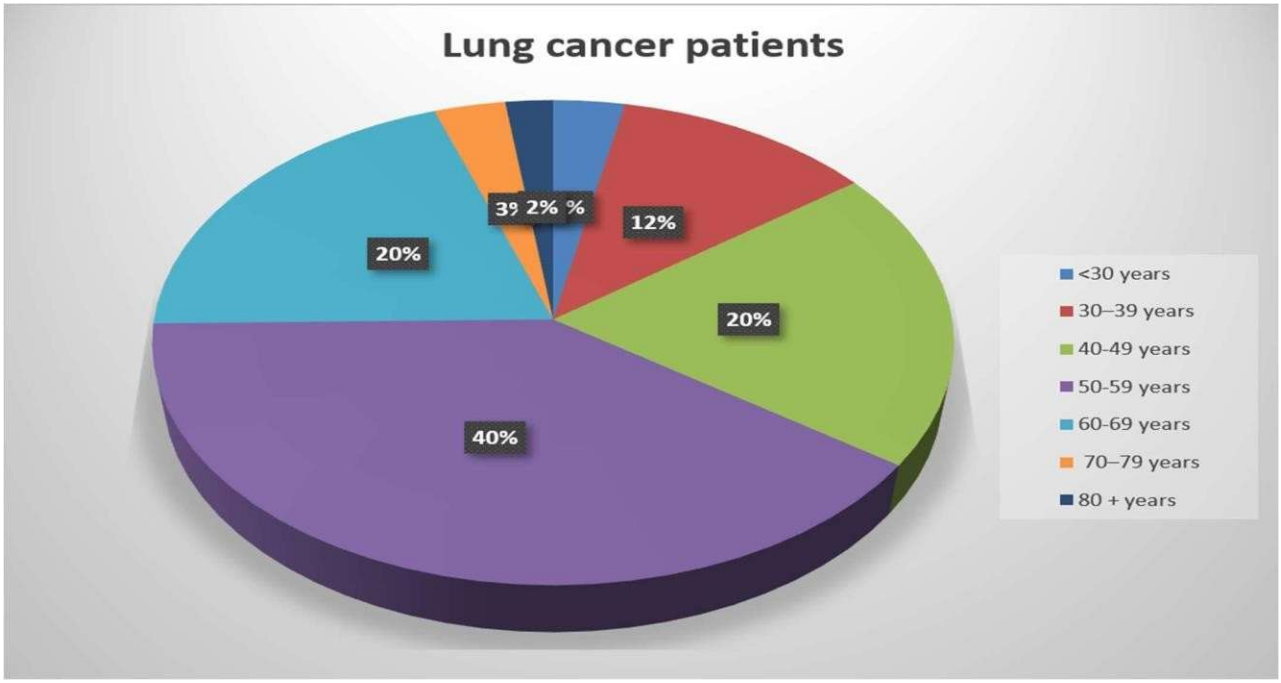

Fig A.III.2 Data for Lung Cancer Patients in India (Age wise)

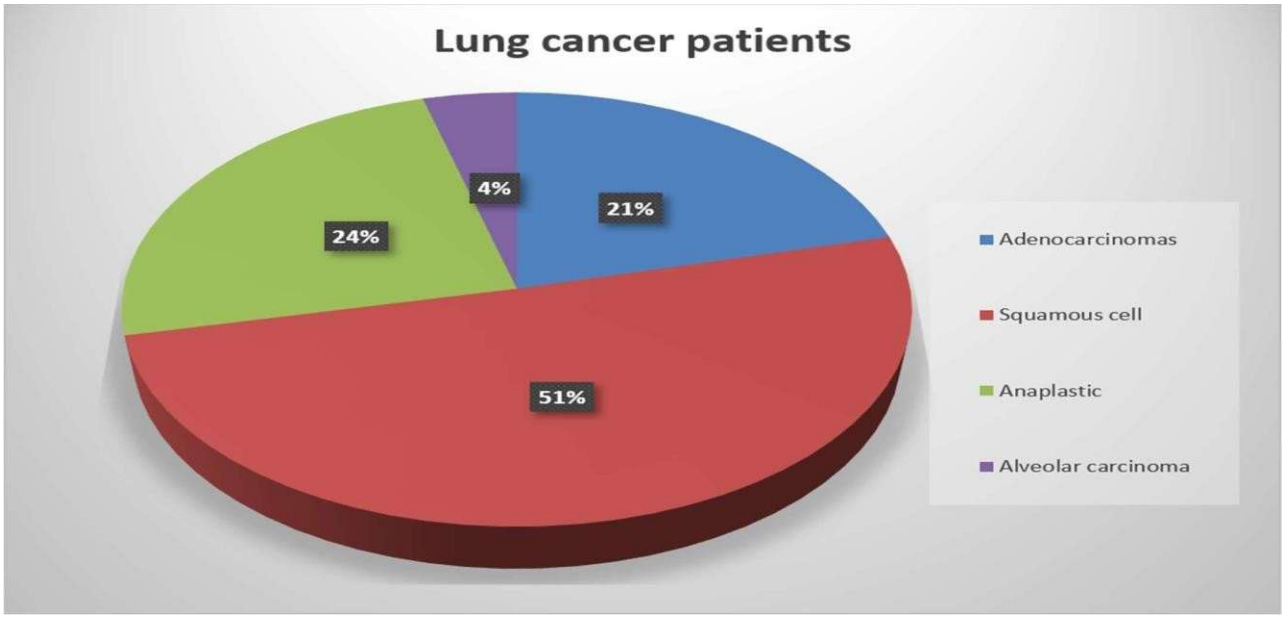

Fig A.III.3 Data for Lung Cancer Variants in India

ACKNOWLEDGEMENT

We have great pleasure in presenting the report on "Computer vision model for lung cancer detection using biopsy images". We extend our heartfelt gratitude to the Head of the Department, Dr. Phiroj Shaikh, for providing invaluable support and encouragement throughout this project. His guidance has been instrumental in shaping the direction of our research. We would like to express our sincere appreciation to the Mentor of the Department, Dr. Amiya Kumar Tripathy, for his insightful feedback and constructive criticism. His expertise has been a guiding light, steering this project towards its successful completion. We are deeply thankful to Ms. Mayura Gavhane, our dedicated Project Coordinator, for her continuous support and coordination. Her assistance has streamlined various aspects of our research and contributed significantly to its organization. A special mention goes to our Project Guide, Ms. Dipti Jadhav, whose expertise and mentorship have been invaluable. Her constant encouragement, technical insights, and unwavering support have been instrumental in the successful execution of this project.

We would also like to thank the library staff for helping us with resources and for using the library for our Mini project meetings. Lastly, we would like to thank our family members and friends, and also all the individuals who have indirectly contributed to this research effort. Your support, whether acknowledged here or not, has been pivotal in this endeavor.

Project Team Members:

1. Aaradhya Deotale T. E. – 12

2. Figo Fernandez. T. E. – 20

3. Jess John T. E. – 32