Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction

PROJECT REPORT

21AD1513- INNOVATION PRACTICES LAB

Submitted by

MOHAN KUMAR B 211422243197

KISHORE K S 211422243163

MOHAN KRISHNA R 211422243196

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

PANIMALAR ENGINEERING COLLEGE, CHENNAI-600123

ANNA UNIVERSITY: CHENNAI-600 025

November, 2024

BONAFIDE CERTIFICATE

Certified that this project report titled "**Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction**" is the bonafide work of **Mohan Kumar B, Kishore K S, Mohan Krishna R**, Register No.: **211422243197, 211422243163, 211422243196,** who carried out the project work under my supervision.

Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

INTERNAL GUIDE
Mr. S. DINESH M.E
Assistant professor,
Department of AI &DS.

HEAD OF THE DEPARTMENT
Dr.S.MALATHI M.E., Ph.D
Professor and Head,
Department of AI & DS.

Certified that the candidate was examined in the Viva-Voce Examination held on ……

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

Lung cancer is one of the most common and deadliest cancers in the world, so early detection for timely treatment is very important. Here we describe the development and evaluation of machine learning models predicting lung cancer status using a combination of clinical measurements including radiologist-computed percent calcification, patient demographics (race, age, gender with smoking status / family history). This paper focuses on different algorithms including Logistic Regression (LR), Decision Trees (DTs), Random Forests (RF) and Support Vector Machines (SVM) as well state of the art deep learning approaches like Convolutional Neural Network (CNN) applied for image-based analysis. The methods applied in this research are key to maximizing prediction accuracy, the ultimate aim of data preprocessing, feature selection and model optimization. We trained the models and assessed their performance on metrics like accuracy, precision, recall, F1 score, and area under the ROC curve. This shows how machine learning and deep learning models can be used to improve lung cancer prediction and early diagnosis through predictive classification techniques. This is especially true when deep-based specificity criteria and social history data are incorporated.

Keywords **:** Lung Cancer**,** Detection**,** Machine Learning**,** Radiologist-computed**,** Convolutional Neural Network**,** Prediction**,** Patient Demographics**,** Deep Learning

ACKNOWLEDGEMENT

I also take this opportunity to thank all the Faculty and Non-Teaching Staff Members of Department of Artificial Intelligence and Data Science for their constant support. Finally I thank each and every one who helped me to complete this project. At the outset we would like to express our gratitude to our beloved respected Chairman, **Dr.Jeppiaar M.A.,Ph.D,** Our beloved correspondent and Secretary **Mr.P.Chinnadurai M.A., M.Phil., Ph.D.,** and our esteemed director for their support.

We would like to express thanks to our Principal, **Dr. K. Mani M.E., Ph.D.,** for having extended his guidance and cooperation.

We would also like to thank our Head of the Department, **Dr.S.Malathi M,E.,Ph.D.,** of Artificial Intelligence and Data Science for her encouragement.

Personally we thank **Mr.S.Dinesh M.E**, Department of Artificial Intelligence and Data Science for the persistent motivation and support for this project, who at all times was the mentor of germination of the project from a small idea.

We express our thanks to the project coordinators **Mrs.V.REKHA M.E** Associate Professor in Department of Artificial Intelligence and Data Science for their Valuable suggestions from time to time at every stage of our project.

Finally, we would like to take this opportunity to thank our family members, friends, and well-wishers who have helped us for the successful completion of our project.

We also take the opportunity to thank all faculty and non-teaching staff members in our department for their timely guidance in completing our project.

MOHAN KUMAR B
KISHORE K S
MOHAN KRISHNA R

TABLE OF CONTENTS

	LIST OF FIGURES	
FIGURE NO.	**TITLE**	**PAGE NO.**
3.1.1	System Architecture	11
3.1.2	Work flow of the model	13
6.1.1	Correlation Heatmap	23
6.1.2	Prediction result	23
6.1.3	Performance comparison of models	24
6.2.1	Stages of lung cancer	25
6.2.2	Confusion matrices	26
6.2.3	Performance comparison of models (Image)	27
6.2.3	Prediction result (Image)	27

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Expansions
AI	Artificial Intelligence
ML	Machine Learning
SVM	Support Vector Machine
DL	Deep Learning
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
LDCT	Low-Dose Computed Tomography
MRI	Magnetic Resonance Imaging
TNM	Tumor, Node, Metastasis
COPD	Chronic obstructive pulmonary disease

CHAPTER 1

INTRODUCTION

1.1 Introduction to Lung Cancer

Lung cancer is among the most prevalent cancers and remains the primary cause of cancer-related mortality across genders. Early detection is vital to enhance survival outcomes, as timely intervention significantly improves the patient's prognosis. However, various factors contribute to delays in diagnosis, making it challenging for healthcare providers to promptly identify patients who require further evaluation [4]. Lung cancer has a higher mortality rate than the combined deaths from breast and colon cancers. This fatal illness develops due to the uncontrolled multiplication of malignant cells in one or both lungs. Early screening and timely diagnosis can enhance the survival rate of lung cancer patients [2]. Through AI & ML, the model will have better generalization ability in more complex and large amounts of data compared to before. Furthermore, symptoms such as coughing up blood, wheezing, and nail clubbing are key signs that may indicate the presence of lung cancer. Along with smoking, aging, and hereditary factors, lifestyle habits like alcohol intake and exposure to polluted air must be taken into account. [7].

How can lung cancer be detected or diagnosed?

There are several techniques such as imaging techniques, biopsy procedures, pathology and molecular testing;

- Low-Dose Computed Tomography (LDCT) : It is a primary screening method for high-risk populations also it provides detailed images of the lungs.

- Chest X-rays : It is less sensitive than LDCT and can be used

in a few cases.

- MRI : It is usually done to evaluate metastasis or brain involvement.

- Bronchoscopy : It is a procedure in which a thin tube is inserted through the nose or mouth into the lungs to collect tissue samples.

- Needle Biopsy : A thin needle is inserted through the chest wall to obtain a sample from a suspicious area.

- Surgical Biopsy : This is removing a part of the lung or lymph nodes for examination if other methods are inconclusive.

- Pathology testing : Samples are examined by a pathologist to determine if it is non-small cell lung cancer or small cell lung cancer and to stage the same using the TNM classification.

1.2 ML and DL models in Lung cancer prediction

ML and DL, together with AI, have been valuable resources for categorization and prediction in healthcare. Analysts can sort out potential patients into elevated, moderate, or minimal-risk categories using predictive algorithms. Critical factors can also be found from complex datasets for more accurate prognosis.

Logistic regression, decision tree, and SVM are some of the applied ML algorithms that can build predictive models to determine one's risk of developing lung cancer based on smoking, family history, and socio-demographic information.

These models can be employed to identify at-risk populations that need to be screened and intervene early. ML techniques can help clinical data analysis to predict survival rates and the prognosis of the patient in response to various treatments, thus personalizing their treatment plans. Medical images, for example, are analyzed for the presence of lung nodules, tumors, or other abnormalities through the aid of ML algorithms:

Feature Extraction: Identifying relevant features from images that correlate with lung cancer presence or risk.

Classification: Classifying images as benign or malignant based on learned patterns.

DL models can discover complex patterns and features automatically from raw image data and, often with much greater accuracy than traditional ML techniques, which requires little feature extraction by hand.

High-End Image Recognition : CNN are particularly very good at recognizing images-in this case, excellent in analyzing radiologic images in detecting lung cancers.

Tumor Segmentation : The DL models can accurately enable the segmentation of tumors in an imaging study. It allows doctors to accurately localize and measure tumor characteristics that are most important for treatment planning and monitoring.

1.2.1 Various ML and DL models

Table 1) Comparison of different ML and Dl models

MODEL	ACCURACY	INTERPRETABILITY	COMPUTATIONAL DEMAND	BEST FOR
Logistic Regression	Moderate	High	Low	Baseline Comparisons
Support Vector Machine (SVM)	High	Moderate	Moderate to High	High Dimensional Data
Random Forest	High	Moderate	Moderate	Structured Clinical Data
Gradient Boosting	Very High	Low to Moderate	High	Large, Structured Data
XGBoost	Very High	Low to Moderate	High	Imbalanced or Large Structured Data
K–Nearest Neighbors (KNN)	Moderate to High	Low	Moderate to High	Small Datasets or Simple Patterns
Naive Bayes	Moderate	High	Low	Text Classification with Strong Independence Assumptions
Convolutional Neural Network (CNN)	Very High	Low	Very High	Image Data (e.g., CT Scans)

1.3 Feature Engineering and data augmentation

1.3.1 Feature Engineering

Feature engineering is the process of transforming raw data into meaningful features that enhance model performance. For the case of lung cancer prediction, this can include the following:

a. Medical Imaging Features :

Texture Analysis: Extract features concerning the texture of lung nodules from CT scans.

Shape Features: Analyze the shape and size of nodules (e.g., volume, roundness).

b. Clinical and Demographic Features

Patient History: Features like age, smoking history, family history of lung carcinoma, occupational exposures.

Comorbidities: These are other health conditions in a patient that may relate to an increased risk of carcinoma or alter treatment (examples: COPD, asthma).

1.3.2 Data Augmentation

Data augmentation refers to the creation of more training images from existing images so as to have more variations of images in the dataset. For lung cancer prediction in the case of radiographic images, it includes

Image Transformations : Rotate, translate, or flip the CT images for increased variability. Changing the brightness, contrast, or the noise level to simulate conditions of other imaging modalities and randomly apply distortions that are elastic and help in the development of an invariance for small distortions.

1.3.3 Integration Methods and benefits

Feature engineering and data augmentation could be integrated through ;

a. Preprocessing pipeline : Set up a preprocessing pipeline that first applies a pipeline of data augmentation on to increase the size of a dataset and variability and the engineered features of images. Then, from these and augmented images, we make a comprehensive feature set.
b. Training of the Model : Train models with the augmented data set, with a big focus on the fact of engineered features. For example, one can use traditional ML algorithms, such as Random Forest or SVM or DL models, e.g., CNN. Using cross-validation check further on whether the model generalizes well to unseen data.
c. Evaluation and Hyperparameter Tuning : Evaluating your model in a validation set so that one can test how feature engineering and augmentation might have affected generalization performance of the network. Hyperparameter tuning is done further to optimize the performance of the model with the new feature set and the augmented data.

Advantages of Integration

- Strong Model : It will be better in performance since the dataset is more diversified.
- Less Overfitting : It helps avoid overfitting, especially when the dataset is very small.
- Improved Interpretability : Engineered features might give insight into which factors contribute most to lung cancer predictions, aiding clinical decisions.

CHAPTER 2

LITERATURE REVIEW

2.1 "Lung Cancer Prediction based on KNN, Logistic Regression, and Random Forest Algorithm"

This study focuses on predicting lung cancer incidence using machine learning algorithms, specifically K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. A Kaggle dataset containing various lifestyle, demographic, and health factors was used to identify correlations and evaluate model performance. Data preprocessing and visualization revealed significant correlations, particularly between lung cancer and factors like allergies, alcohol consumption, and lifestyle habits. Models were evaluated using accuracy, F1-score, and weighted averages. The Random Forest model outperformed others with an accuracy improvement from 0.96 to 0.98 after tuning with 75 neural networks. The study underscores the benefits of using machine learning to decision trees. It highlights the superior accuracy of the Random Forest algorithm for lung cancer prediction and suggests future integration of other models such as SVM and assess cancer risk efficiently, with potential applications in public health and preventative measures. [6]

Author : Yunzhe Liao

Year : 2024

2.2 "The Efficacy of Machine Learning Models in Lung Cancer Risk Prediction with Explainability"

This study evaluates multiple machine learning models (SVM, KNN, Decision Tree, and Random Forest) for predicting lung cancer risk based on a dataset of risk factors like smoking, genetic history, and occupational hazards. It focuses on the explainability of each model, utilizing techniques like LIME and SHAP to clarify

model predictions, thereby addressing the common "black-box" issue in ML-based medical applications. The study achieves high accuracy across models, with Random Forest reaching near-perfect results, and highlights the role of explainable AI in building trust among healthcare users. Provides a comparative analysis of model accuracy, emphasizes the importance of model transparency in medical settings, and demonstrates explainability techniques for model predictions. [8]

Authors : Refat Khan Pathan, Israt Jahan Shorna , Mayeen Uddin Khandaker, Md. Sayem Hossain, Huda I. Almohammed, Zuhal Y. Hamd

Year : 2024

2.3 "Machine Learning Techniques for Lung Cancer Risk Prediction Using Text Dataset"

The study presents a machine learning approach for predicting lung cancer risk, focusing on analyzing text datasets (specifically electronic health records) containing demographic, clinical, and historical medical data. Unlike traditional cancer screening methods, which often rely on imaging, this study explores various machine learning models (e.g., Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Naive Bayes) to predict lung cancer risk by identifying correlations within non-imaging data.

The article reviews multiple machine learning and deep learning models applied to lung cancer detection and prognosis, including deep learning's advantages for accuracy over traditional models in similar cancer prediction tasks. This work aligns with ongoing research in non-imaging-based lung cancer diagnostics, emphasizing the need for early detection tools using accessible data sources, such as patient records, to improve healthcare outcomes. [11]

Authors : Kumar Mohan, Bhraguram Thayyil

Year : 2023

2.4 “Deep Learning Techniques for Lung Cancer Recognition”

This paper examines deep learning models for lung cancer recognition, addressing the limitations in conventional screening due to the size, shape, and texture variability of lung nodules. Prior studies in lung cancer recognition have often relied on manual screening methods and image processing with traditional machine learning algorithms like Sequential Flood Feature Selection and Genetic Algorithms, which optimize feature selection but present limitations with high false positives and radiation risks in methods like CT. The paper evaluates these models on an extensive dataset of CT images, focusing on VGG-16, which achieved a 95% accuracy rate, outperforming other models. This research builds on related works, such as those that applied DL methods like U-Net, ResNet, and computer-aided design (CAD) frameworks, which aid in extracting high-level features from medical images for improved diagnosis. CAD frameworks have advanced accuracy but require improvements in sensitivity and efficiency to reduce processing time. This paper contributes to the field by demonstrating that ensemble approaches and transfer learning, particularly with VGG-16, can significantly enhance lung cancer detection accuracy. The study highlights the potential of DL in overcoming challenges in lung cancer diagnosis but also acknowledges ongoing obstacles, including the need for privacy protection, large-scale clinical validation, and ethical considerations. [13]

Authors : Suseela Triveni Vemula, Maddukuri Sreevani, Perepi Rajarajeswari, Kumbham Bhargavi, Joao Manuel R. S. Tavares, Sampath Alankritha
Year : 2024

2.5 “Augmented Lung Cancer Prediction: Leveraging Convolutional Neural Networks and Grey Wolf Optimization Algorithm”

This paper proposes a hybrid approach for lung cancer prediction that combines Convolutional Neural Networks (CNN) with the Grey Wolf Optimization

Algorithm (GWOA) to improve diagnostic accuracy and reduce false negatives. GWOA is implemented here for hyperparameter tuning and feature selection, enhancing the performance of CNN by identifying the most relevant features in CT scan images. This paper builds upon these advancements, demonstrating that CNNs combined with GWOA can achieve an average accuracy of 96% with a false-negative rate of 0.0237, significantly enhancing diagnostic precision. This study contributes to the field by showcasing the potential of using optimization algorithms in tandem with CNNs for improved lung cancer diagnosis, with results that surpass traditional methods and standalone machine learning approaches. [16]

Authors : Teresa Kwamboka Abuya, Wangari Catherine Waithera, Cheruiyot Wilson Kipruto
Year : 2024

2.6 "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images"

This comprehensive review examines state-of-the-art deep learning approaches for lung cancer screening and diagnosis via CT images. It categorizes techniques based on tasks such as classification and segmentation, highlighting CNNs, U-Net, and 3D CNN architectures as particularly effective. Deep learning enables precise automated detection, analyzing features like nodule size, shape, and texture.It provides a detailed comparative analysis of current deep learning models, focusing on the advantages and challenges within lung cancer diagnosis. Discusses common model limitations such as high computational needs, and the reliance on extensive annotated datasets. [23]

Authors :Mohammad A.Thanoon , Mohd Asyraf Zulkifley, Muhammad Ammirrul Atiqi Mohd Zainuri, Siti Raihanah Abdani
Year : 2023

CHAPTER 3

SYSTEM DESIGN

3.1 System Architecture

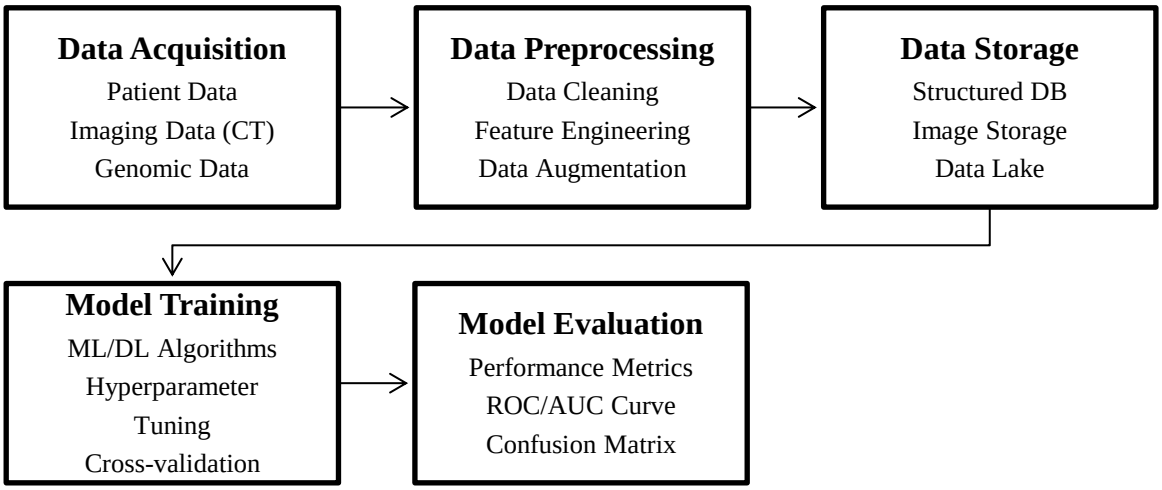

Fig.3.1.1 System architecture

1. Data Acquisition
   - Patient data: demographic and medical background information and clinical variables
   - Imaging data: CT scans or radiographs related to the lung and lung health.
   - Genomic data: genetic mutation or markers pertinent to the case of lung cancer

2. Preprocessing data
   - Data cleaning : Remove duplicates, handle missing values, and ensures data quality.
   - Feature Engineering: Extract and create meaningful features from raw data like texture features from images or clinical variables.
   - Data Augmentation: Create additional training samples for imaging data,

thereby increasing diversity and avoiding overfitting.

3. Data Storage
    - Structuring the Database: In this approach, patient data, features, and labels are stored in a relational database such as PostgreSQL or MySQL.
    - Storage for Image : File system or object storage, for instance AWS S3.
    - Data Lake: Raw data as extracted, unprocessed in ready for further analysis or processing.

4. Train Your Model
    - ML & DL Algorithms : We can utilize various ML and DL algorithms.
    - Hyperparameter Tuning : Optimize model parameters using techniques like grid search or random search.
    - Cross-validation: Assess model performance on different subsets of the data to ensure robustness.

5. Model Evaluation
    - Performance Metrics: Accuracy, precision, recall, F1 score, etc.
    - ROC/AUC Curve: To see the trade-off between the true positive rate and false positive rate.
    - Confusion Matrix: Number of true positives, false positives, true negatives, and false negatives.

3.2 Flow of the model

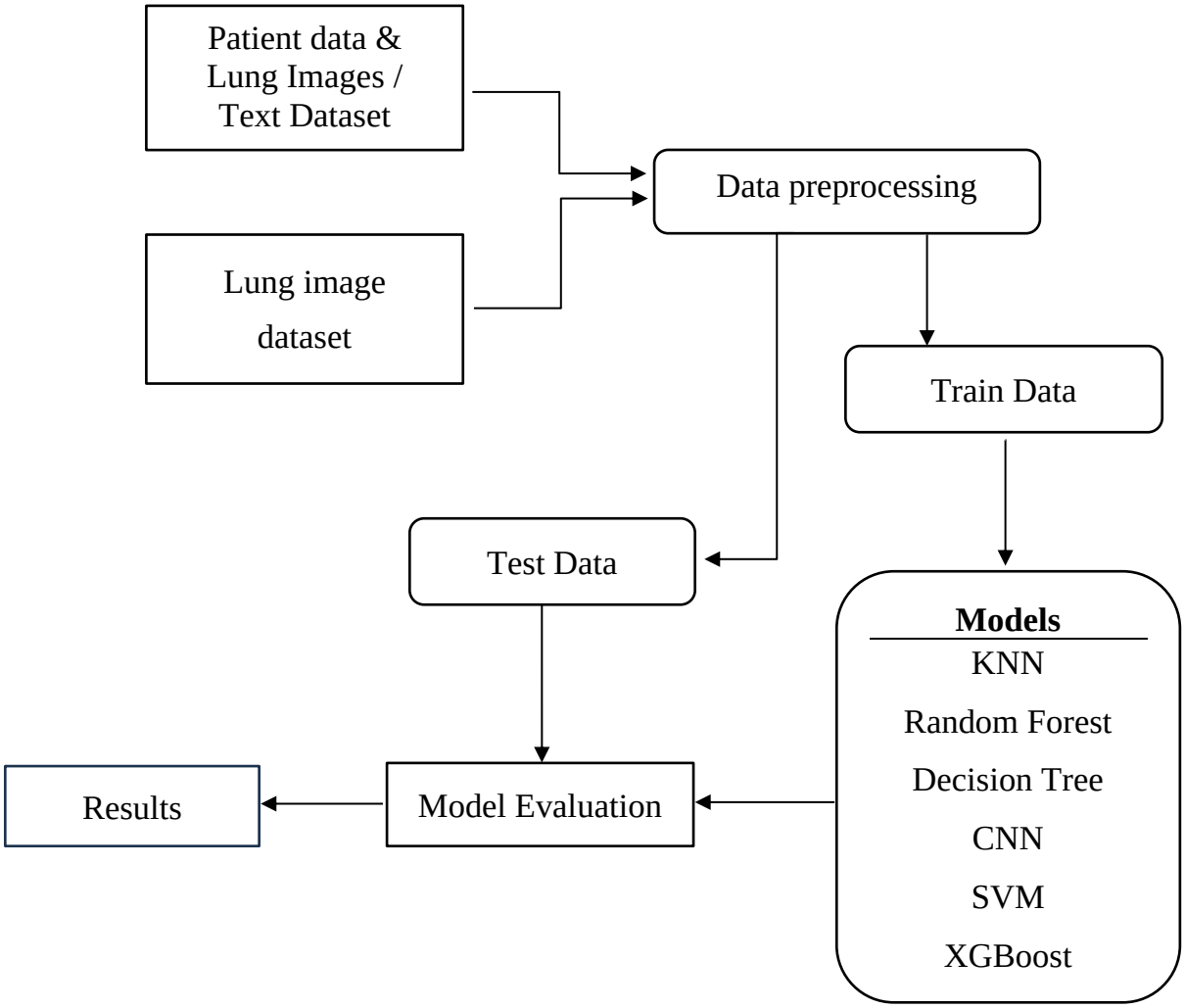

Fig. 3.2.1 Work flow of the model

This is the work flow of the model as the images/text inputs are collected then preprocessed for Train set and Test. The dataset is divided into training and testing sets, with the training set serving as the basis for model training and the testing set as the basis for model testing. Then the models are evaluated and assessed for their performance.

CHAPTER 4

PROJECT MODULES

The project consists of Four modules. They are as follows,

1. Building the prediction model
2. Training and evaluation
3. Comparison of the models
4. Prediction of lung cancer

4.1 Building the prediction model

It makes use of a deep learning algorithm, especially CNN, for detection and diagnosis of cancer by the nodules of cancerous cells found in lungs using CT images. The model is based on the Sequential class where layers are piled up in a linear sequence. The CNN architecture consists of a serial model that starts from the use of three layers of convolutional layers, employing progressively increasing numbers of filters of 32, 64, and 128, using ReLU activation functions intended to inject non-linearity, with max pooling in place in order to down-sample the feature maps spatially to reduce their size. In the final stage, this is followed by a flatten layer that unfolds the two-dimensional features to be represented as a vector of 1-D elements to make them ready to feed denser layers. Including a fully connected dense layer of 512 units Dropout layer set to 50% to avoid overfitting Final layer of the network: using softmax activation since three classes need to be predicted - benign, malignant, and normal instantiates the model with room for further training and evaluation. This CNN model is structured to effectively classify lung cancer images by learning relevant features through convolutional layers, capturing complex patterns, and employing dropout layers to enhance generalization, which is a basic step towards

developing a robust model for lung cancer detection through image classification techniques.

4.2 Training and evaluation

The model is trained over 10 epochs with a batch size of 32 on the prepared training data. It uses a validation split of 20% and trains the model to evaluate its performance over unseen validation data during the process. In addition, the training history with the accuracy and loss value at each epoch is kept for later analysis. Finally, after training, the model is adapted to be used as a feature extractor. Using the Keras Functional API, a new model is defined that outputs the activations from the second-to-last layer of the CNN, representing high-level features for each input image. Vectors of features are created both for the training and testing sets. These can further be used in downstream tasks, such as clustering, or as an input to some other machine learning models for further analysis. The model's performance can be looked upon by the history plot function that uses Plotly to plot the training as well as validation accuracy as well as loss over the epochs. Two subplots were created, one for the case of accuracy and another one for loss. It contains scatter plots for both training and validation values to show a trend across epochs; therefore, signs of overfitting or underfitting can be identified. The performance of the trained model on the test set is evaluated using the evaluate function by returning the test loss and accuracy. This step directly tests the generalization ability of the model on unseen data. The feature extraction model is applied again to both the training and test data, where it produces feature vectors from the second-to-last dense layer. These are the high-level embeddings, which provide a compact and informative representation of images, and can be useful for further analyses or to build secondary models.

4.3 Comparison of the models

This describes the performance summary and evaluation of several machine learning models trained to classify images into three classes. The performance of the

models can be measured through accuracy metrics, confusion matrices, as well as a comparative bar chart showing several metrics of interest. Accuracy scores from each model are printed with this code for a precise overview of performance. These models are assessed in this paper by accuracy, F1 score, precision, recall, True Positive Rate, and False Negative Rate. This plot is a good insight of both models that capture all correct classifications as compared to false classifications. Now for gaining insight into which part the models are really failing and to what extent it happens, confusion matrices of the same are plotted here as well. In that case, for each of these three plots, the number of True Positive, True Negative, False Positive, False Negative distributions for all Benign, Malignant, Normal classes can be determined. Misclassifications can be identified by simply visually comparing confusion matrices; they highlight any patterns whereby the models might struggle with certain classes. Then goes further into analysing the models by creating a dataframe which would store the metrics for every model. This includes accuracy, F1 score, precision, recall, TPR, and FNR. These metrics are chosen because they collectively assess model performance from multiple angles, including balance between precision and recall (F1 score), prediction quality (precision and recall), and identification of positive cases (TPR) and negatives (FNR). A bar chart is constructed to visually present these metrics for each model, one colour for each. This chart will make you easily compare the models side-by-side and therefore notice that which model has the highest values for each of the metrics. For instance, it shows that high TPR and low FNR means how robust the model will be while identifying the cases as positive, and with high accuracy and precision indicates the capability of the model toward making the right calls. The bar chart will effortlessly highlight the best models that will further lead to optimizing and selection of the right model.

4.4 Prediction of lung cancer

The prediction is done by processing images by first checking if the input is a directory or a single image file. Each image is converted to grayscale, resized to standard dimensions, and normalized to a specific pixel intensity range. It is then reshaped to fit the expected input format for the CNN model. After preprocessing, the image is fed into a pre-trained model to obtain a prediction. The class with the highest probability is selected as the predicted label. This program outputs a single integer (i.e. 0 or 1 or 2) representing the corresponding index of the predicted class. If the input is a directory, the code iterates through each image, applying the same preprocessing and prediction steps, and prints the classification result for each image.

CHAPTER 5

SYSTEM REQUIREMENTS

5.1 Introduction

This chapter will involve technology used, the hardware and the software requirements of the project. The demands of processing large datasets along with execution of more complex machine learning algorithms require high-performance hardware from the system. A suitable system with adequate resources in hardware and software can run and execute the code. It generally requires a modern processor, a suitable GPU that provides support for CUDA, mainly for deep learning acceleration, on the hardware side. On the software side, it would need to run Python version 3.8 or higher with the required packages installed. For smoother processing, an SSD and IDE are recommended to ensure smooth execution of code and data handling.

5.2 Requirements

5.2.1 Hardware Requirements

The hardware requirements for the prediction and classification of lung cancer are critical to ensuring its efficiency, reliability, and capability to handle demanding computational tasks. Below are the required specifications;

- **Processor**:
  - **Intel i7/i9** or **AMD Ryzen 7/9 (Recommended) :** These processors provide multiple cores and threads, significantly enhancing parallel processing capabilities, crucial for running complex algorithms and analyses.
  - **Intel i5 or AMD Ryzen 5 (Minimum) :** These processors efficiently manage demanding workloads but is slower and has less cores and threads.

- **GPU**:
  - **NVIDIA RTX 2050 or higher**: Accelerating training of the DL models requires a powerful GPU because these GPUs contains the parallel-processing capability that significantly accelerates the training of deep-learning models for example with tasks such as object detection and image classification.

- **RAM**:
  - **Minimum of 16GB**: This is the baseline capacity that the system can handle standard workloads and smaller datasets without performance degradation.
  - **32GB Recommended**: For optimal performance, especially when working with large datasets or running multiple applications simultaneously, 32GB of RAM is recommended.

- **Storage**:
  - **SSD (Solid State Drive)**: Storage with at least **1TB capacity** is recommended for fast data access and the storage of video footage, images, and processed outputs. SSDs greatly enhance system responsiveness and reduce loading times, making them ideal for handling the large volumes of data typical in forensic investigations. And the minimum storage space is **512GB**.

5.2.2 Software Requirements

The software requirements for the prediction and classification of lung cancer are essential to ensure effective development and operation of the system. Below are the key software components necessary to support the various functionalities of the system:

- **Operating System**:
    - **Windows 10/11 or Linux-based OS (Ubuntu preferred)**: The choice of operating system plays a crucial role in supporting the development and deployment of forensic analysis applications. Windows provides a user-friendly environment, while Linux, particularly Ubuntu, is favoured for its stability, flexibility, and robust support for open-source tools, making it an excellent choice for running machine learning frameworks and handling server-side applications.

- **Programming Language** :
    - **Python**: This language is widely used for its simplicity and versatility, making it ideal for implementing machine learning models and conducting data analysis. Its rich ecosystem of libraries facilitates rapid development and experimentation.
    - **Libraries** :

      Key libraries for ML, DL and data processing include:
        - **TensorFlow** : An open-source library developed by Google that provides extensive support for deep learning applications, including neural networks for image and video analysis.
        - **OpenCV** : OpenCV can be used for reading and manipulating images. It is particularly for performing image preprocessing.
        - **NumPy**: For numerical operations.
        - **Pandas**: For data handling and table creation.
        - **Matplotlib & Seaborn**: For plotting and visualizing data and metrics.
        - **Plotly**: For interactive and advanced plotting (used in accuracy and loss plots).
        - **Scikit-learn**: For classical machine learning models, confusion matrix, and evaluation metrics.

- **Development Environment**:
    - **Integrated Development Environments (IDEs)**: Effective coding, testing, and debugging require robust development environments. Recommended IDEs include:
        - **Jupyter Notebook**: This interactive environment is particularly well-suited for data analysis and visualization, allowing developers to document their code alongside rich media and visualizations.
        - **Visual Studio Code:** A versatile code editor that supports multiple programming languages, including Python. It has excellent extension support for Python development.

CHAPTER 6
CONCLUSION & RESULTS

This chapter provides a comprehensive summary of the project "**Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction**", highlighting key findings, achievements, and insights derived from system testing and performance analysis. This chapter includes results obtained from the prediction and classification of the lung cancer and the models using both structured and image data.

6.1 RESULTS

6.1.1 Results from structured data

The dataset used for this prediction contains 276 entries and 16 columns. The dataset includes 14 integer datatype columns (14 columns of type int64) and 2 object columns. The dataset was suitable for binary classification tasks, based on symptoms and lifestyle factors that causes lung cancer.

The Fig 6.1.1 is called a correlation heatmap, it visually represents the correlation between different features in a dataset. Correlation does not imply causation while the heatmap only shows relationships, it doesn't prove that one feature directly causes another.

The features with strong positive correlations are likely to influence each other, the color of each cell indicates the strength and direction of the correlation between two features.

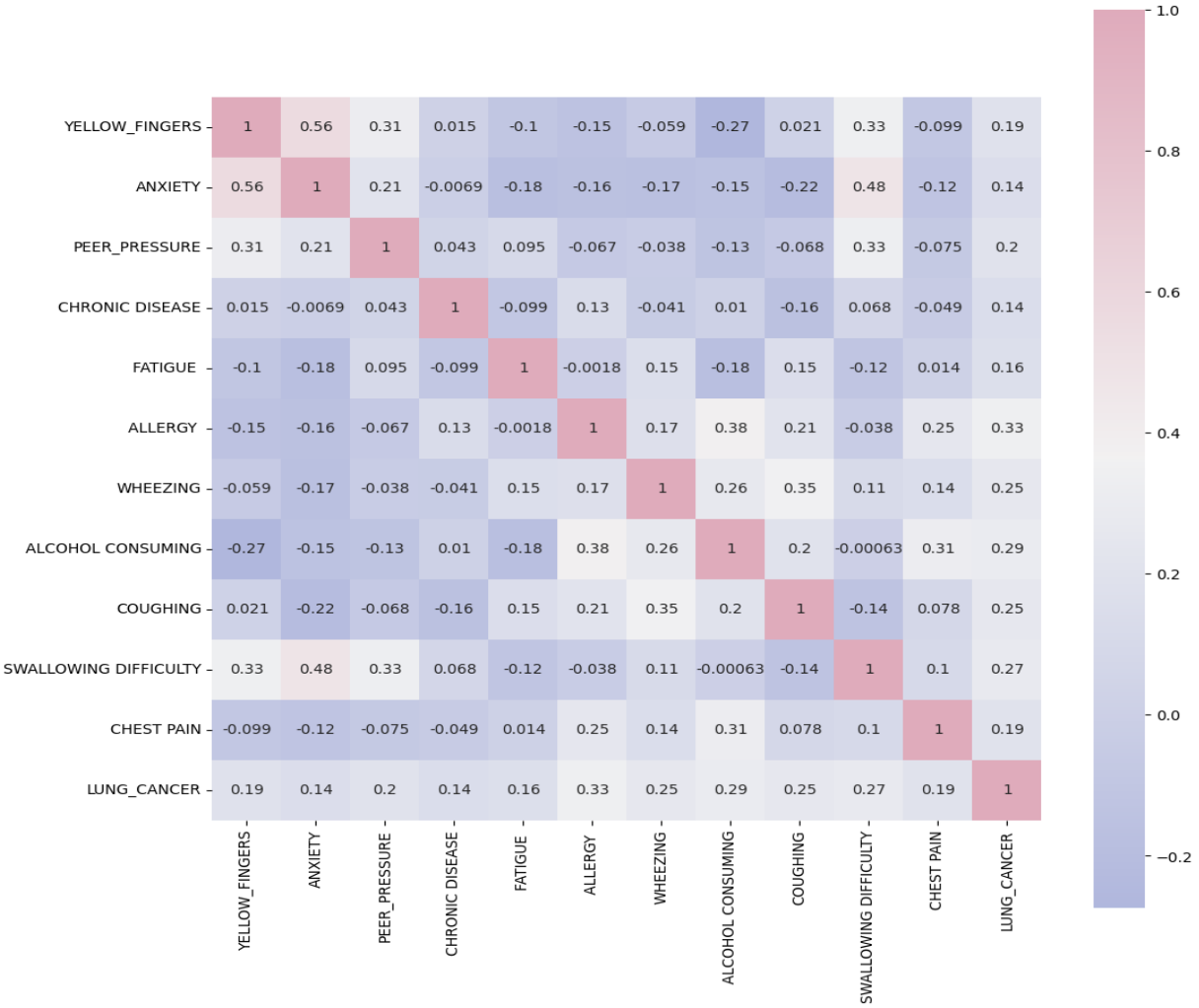

Fig 6.1.1) Correlation Heatmap

```
Please enter the following information:
Peer Pressure (1 for Yes, 0 for No): 0
Yellow Fingers (1 for Yes, 0 for No): 1
Anxiety (1 for Yes, 0 for No): 1
Chronic Disease (1 for Yes, 0 for No): 0
Fatigue (1 for Yes, 0 for No): 1
Allergy (1 for Yes, 0 for No): 0
Wheezing (1 for Yes, 0 for No): 0
Alcohol Consuming (1 for Yes, 0 for No): 1
Coughing (1 for Yes, 0 for No): 1
Swallowing Difficulty (1 for Yes, 0 for No): 0
Chest Pain (1 for Yes, 0 for No): 1
Prediction: Positive for Lung Cancer
```

Fig 6.1.2) Prediction result

The Fig 6.1.2 represents the prediction output of the model, it represents if the patient has lung cancer or not by getting the input information from the user and predicts if the person has lung cancer or not.

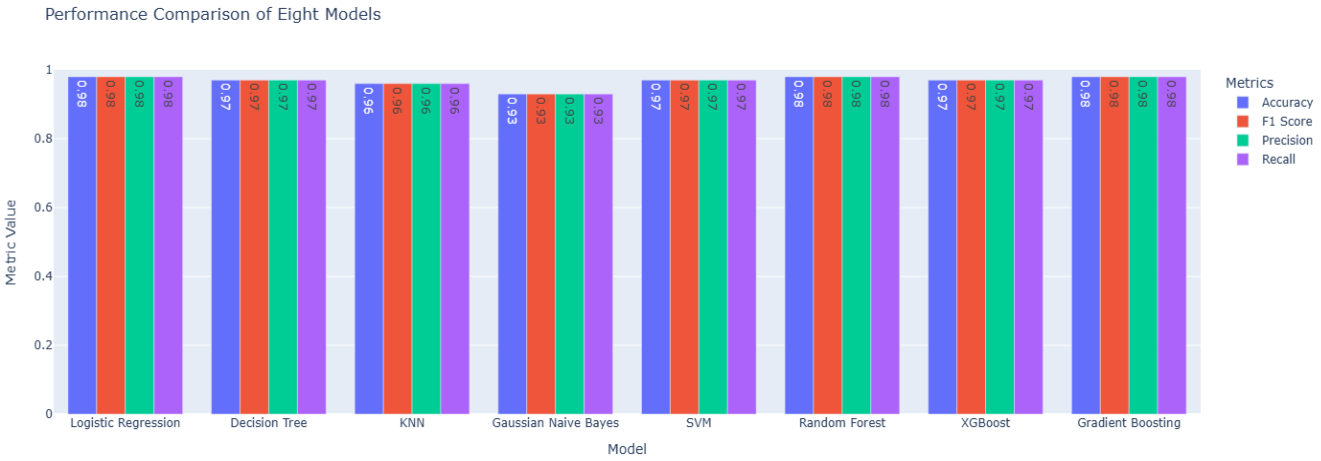

Fig 6.1.3) Performance comparison of models

The Fig 6.1.3 represents the performance of each model, the metrics includes accuracy, f1-score, precision, recall.

6.2 Results from image data

The Fig 6.2.1 represents the data present in the dataset and is used for our prediction model to predict the lung cancer. It consists of three cases “Benign, Malignant and Normal” where each state represents stages of the lung cancer.

- Benign tumors grow slowly and have distinct, well-defined boundaries these do not pose a life-threatening risk but may still require monitoring or surgical removal if they cause complications or symptoms.
- Malignant tumors are cancerous and represent true lung cancer. These tumors have the potential to invade surrounding tissues and spread to other parts of the body, often to lymph nodes, liver, or brain. Malignant tumors grow rapidly and are invasive and can have irregular or poorly defined edges.

- Normal lung tissue indicates the absence of any tumor, benign or malignant. This tissue is free from abnormal growths or cells associated with lung cancer.

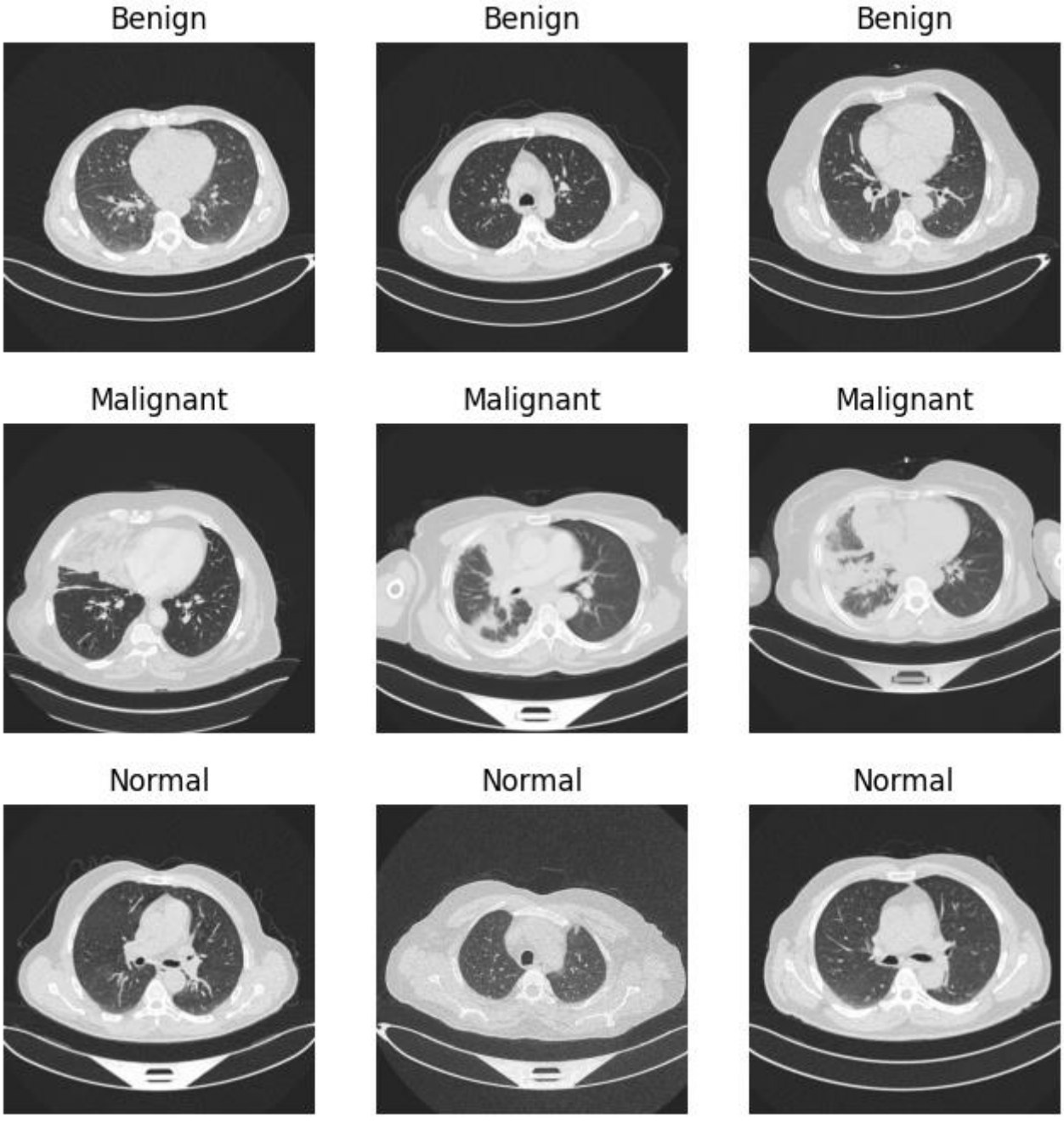

Fig 6.2.1) Stages of lung cancer

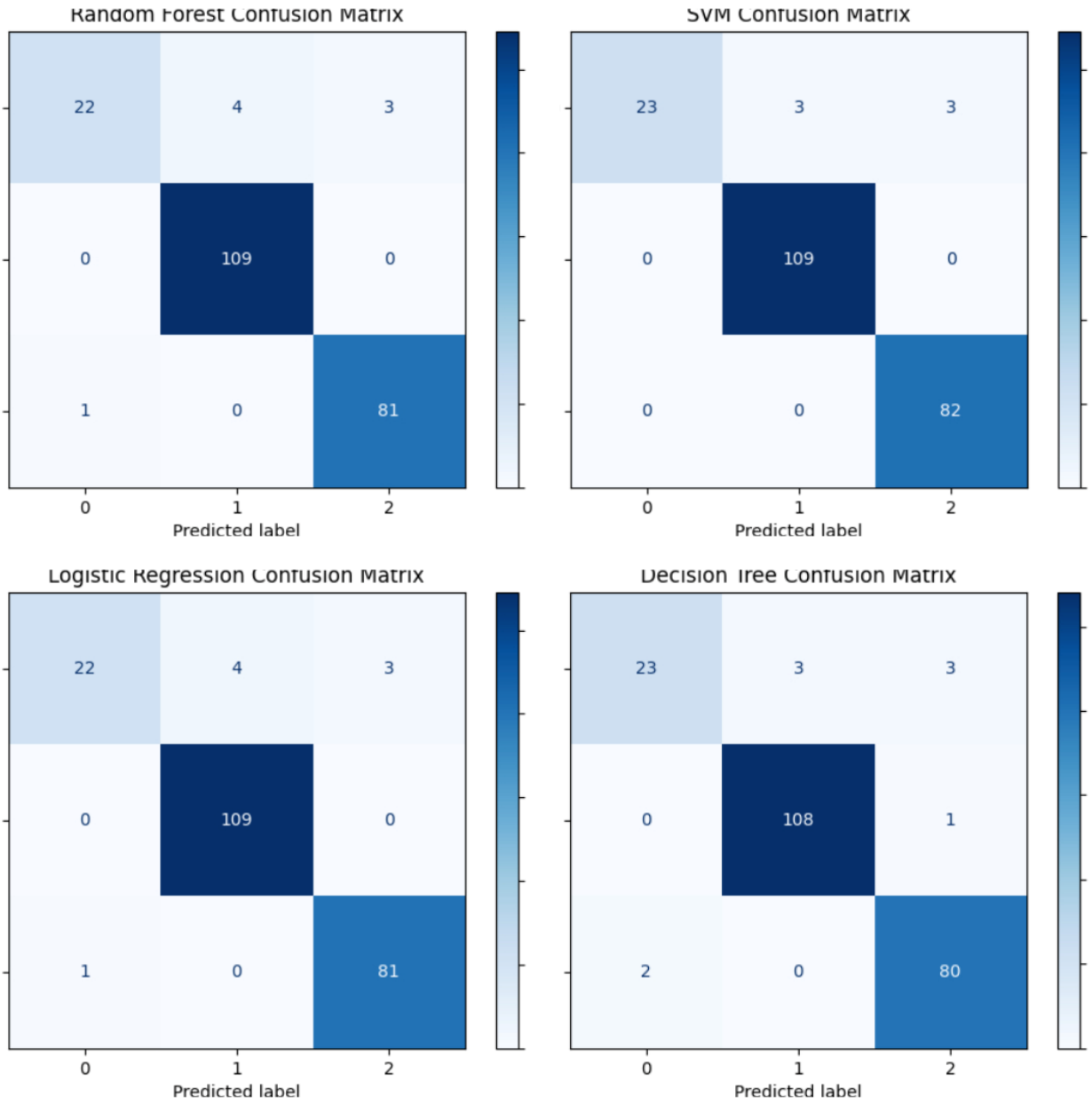

Fig 6.2.2) Confusion matrices

The above Fig 6.2.2 represents the confusion matrices of each models used for the prediction of the cancer using images.

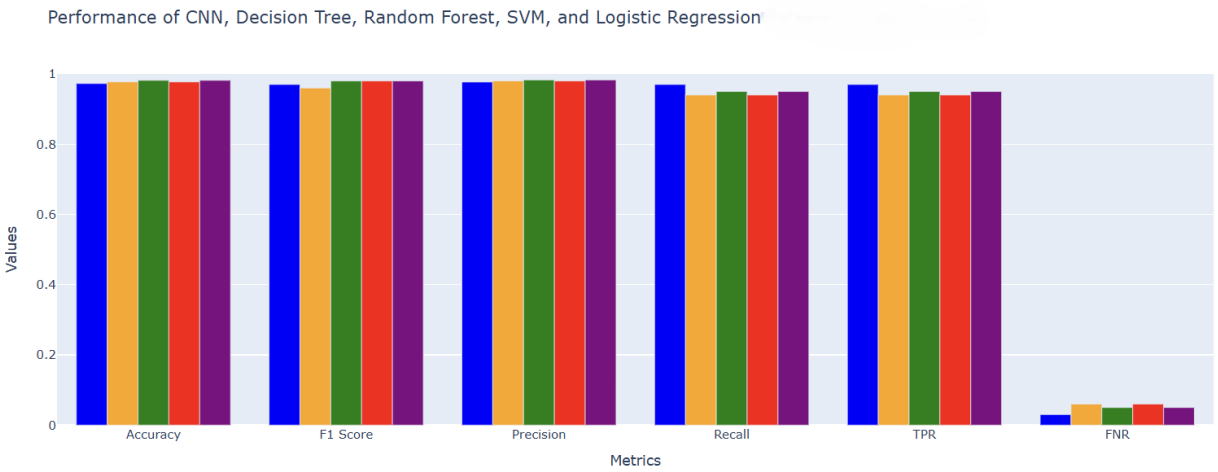

Fig 6.2.3) Performance comparison of the models (image)

The Fig 6.2.3 represents the performance of each of the model used it includes accuracy, f1-score, precision, recall, true positive rate and false negative rate.

```
image_path = "/content/demo 3.jpg"
# Use the predict_class function we defined e
predicted_class = predict_class(image_path)

print("Predicted class:", predicted_class)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 119ms/step
Predicted class: 0
```

```
image_path = "/content/demo 1.jpg"
# Use the predict_class function we defined ear
predicted_class = predict_class(image_path)

print("Predicted class:", predicted_class)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 257ms/step
Predicted class: 2
```

Fig 6.2.4) Prediction result (image)

The Fig 6.2.4 specifies the output that the model has predicted through the input image and the prediction output classifies the image belongs to which class and returns the class index of the corresponding class.

6.2 CONCLUSION

This study develops a lung cancer prediction model using a variety of machine learning techniques, such as CNN, decision trees, random forests, SVM and logistic regression. Major findings from this study concerning the classification and early lung cancer detection will be discussed at length. Overall, thorough analysis and comparison of different models will reflect the effectiveness of machine learning in promoting more accurate and efficient diagnostic tools.

The CNN achieved good performance and high precision; it also acquired deeper features from lung images. It is helpful especially in such image-based classifications and detection of growth of cancerous nodules in lung scans. The traditional models like decision tree and random forests also offered high accuracies. The evaluation metrics were accuracy, precision, recall, and F1-score. These metrics brought out the relative strengths and weaknesses of each model to inform the selection of the most appropriate method for clinical application. Additionally, confusion matrices were provided to give visual insights into model performance, thereby facilitating the identification of specific areas for improvement.

In general, the project indicates the necessity of applying different machine learning methods to predict lung cancer because it allows multiple model validations and supports the production of more reliable diagnostic tools.

REFERENCES

[1] S. U. Krishna, A. N. B. Lakshman, T. Archana, K. Raja, and M. Ayyadurai, "Lung Cancer prediction and classification using Decision Tree and VGG16 convolutional neural networks," *Open Biomed. Eng. J.*, vol. 18, no. 1, 2024.

[2] T. K. Abuya, W. C. Waithera, and C. W. Kipruto, "Augmented lung cancer prediction: Leveraging convolutional neural networks and grey wolf optimization algorithm," *OAlib*, vol. 11, no. 04, pp. 1–25, 2024.

[3] D. Li, G. Li, S. Li, and A. Bang, "Classification prediction of lung cancer based on machine learning method," *Int. J. Healthc. Inf. Syst. Inform.*, vol. 19, no. 1, pp. 1–12, 2023.

[4] K. H. Rubin *et al.*, "Developing and validating a lung cancer risk prediction model: A nationwide population-based study," *Cancers (Basel)*, vol. 15, no. 2, p. 487, 2023.

[5] M. Rhifky Wayahdi and F. Ruziq, "KNN and XGBoost algorithms for lung cancer prediction," *JoSTec*, vol. 4, no. 1, pp. 179–186, 2022.

[6] Y. Liao, "Lung cancer prediction based on KNN, logistic regression, and random forest algorithm," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 280–287, 2024.

[7] X. Li, "Lung cancer risk prediction and feature importance analysis with machine learning algorithm," *Applied and Computational Engineering*, vol. 19, no. 1, pp. 205–210, 2023.

[8] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd, "The efficacy of machine learning models in lung cancer risk prediction with explainability," *PLoS One*, vol. 19, no. 6, p. e0305035, 2024.

[9] F. A. Altuhaifa, K. T. Win, and G. Su, "Predicting lung cancer survival based on clinical data using machine learning: A review," *Comput. Biol. Med.*, vol. 165, no. 107338, p. 107338, 2023.

[10] V. Rajasekar, M. P. Vaishnnave, S. Premkumar, V. Sarveshwaran, and V. Rangaraaj, “Lung cancer disease prediction with CT scan and histopathological images feature analysis using deep learning techniques,” *Results Eng.*, vol. 18, no. 101111, p. 101111, 2023.

[11] K. Mohan and B. Thayyil, “Machine learning techniques for lung cancer risk prediction using text dataset,” *International Journal of Data Informatics and Intelligent Computing*, vol. 2, no. 3, pp. 47–56, 2023.

[12] S. T. Vemula, M. Sreevani, P. Rajarajeswari, K. Bhargavi, J. M. R. S. Tavares, and S. Alankritha, “Deep learning techniques for lung cancer recognition,” *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, pp. 14916–14922, 2024.

[13] Mugdho, Aka Mohammad Bhuiyan, Md. Jawad Hossain Rafin, Tawsif Mustasin Amit, Adib Muhammad, Ed., *A comparative study of lung cancer prediction using deep learning*. Brac University, 2022.

[14] Lithesh Gadikota Venkata Yuva Naga Sai Nunna Sai Teja Tirumani Venkata Vara Prasad Padyala, Ed., *Lung Cancer Prediction by Using Deep Learning method CNN*, no. https://doi.org/10.21203/rs.3.rs-2614821/v1. Research Square, 2023.

[15] S. Al Rumhi Raza Hasan Saqib Hussain Jitendra Pandey, Ed., *Lung Cancer Prediction Using Machine Learning Techniques*, no. 2167–1907,https://www.jsr.org/index.php/path/article/view/2233. Jounal of Student Research, 2023.

[16] E. Dritsas and M. Trigka, “Lung cancer risk prediction with machine learning models,” *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 139, 2022.

[17] M. A. Thanoon, M. A. Zulkifley, M. A. A. Mohd Zainuri, and S. R. Abdani, “A review of deep learning techniques for lung cancer screening and diagnosis based on CT images,” *Diagnostics (Basel)*, vol. 13, no. 16, 2023.

[18] T. Kadir and F. Gleeson, “Lung cancer prediction using machine learning and advanced imaging techniques,” *Transl. Lung Cancer Res.*, vol. 7, no. 3, pp. 304–312, 2018.

[19] Ö. Çelik, “A research on machine learning methods and its applications,” *J. Educ. Technol. Online Learn.*, vol. 1, no. 3, pp. 25–40, 2018.

[20] S. P. Maurya, P. S. Sisodia, R. Mishra, and D. P. Singh, “Performance of machine learning algorithms for lung cancer prediction: a comparative approach,” *Sci. Rep.*, vol. 14, no. 1, p. 18562, 2024.