Leveraging Genetic Data for Lung Disease Prediction using Machine Learning Algorithms

22BIO211 – Intelligence of Biological Systems - II

TEAM – 12

CH.SC.U4AIE23007 - B V S HEMANTH
CH.SC.U4AIE23014 - FAIZ RAHAMAN
CH.SC.U4AIE23015 - SRIMAAN GAJJELLI
CH.SC.U4AIE23019 - HARINDERAN T
CH.SC.U4AIE23049 - SABBU DITESHWAR
CH.SC.U4AIE23064 - RAJ DHANUSH

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

TEAM-12

INTRODUCTION

22BIO211

Lung diseases such as COPD, lung cancer, and pulmonary fibrosis pose a major global health challenge, leading to millions of deaths each year. These diseases result from complex interactions between genetic and environmental factors. This research proposes leveraging machine learning (ML) and artificial intelligence (AI) to enhance the prediction of the lung cancer cells.

TEAM-12

PROBLEM STATEMENT

22BIO211

In this research, we aim to **compare different Machine Learning models and analysing the accuracy of these models in the detection of Lung Cancer** . Our work could help researchers and pathologists in identifying and analyzing the cancer cells in a effective manner.

TEAM-12

WORKFLOW OF PROJECT

22BIO211

PREVIOUS IDEOLOGIES:

- Traditional diagnostic methods relied on imaging (CT scans, X-rays) and visible symptoms.
- Limited ability to detect lung diseases at early stages.

PROPOSED METHOD:

- Machine learning-based approach using **genetic data and AI models**.
- **Feature selection & model optimization** for high-accuracy disease prediction.

TEAM-12

ARCHITECTURE DIAGRAM

22BIO211

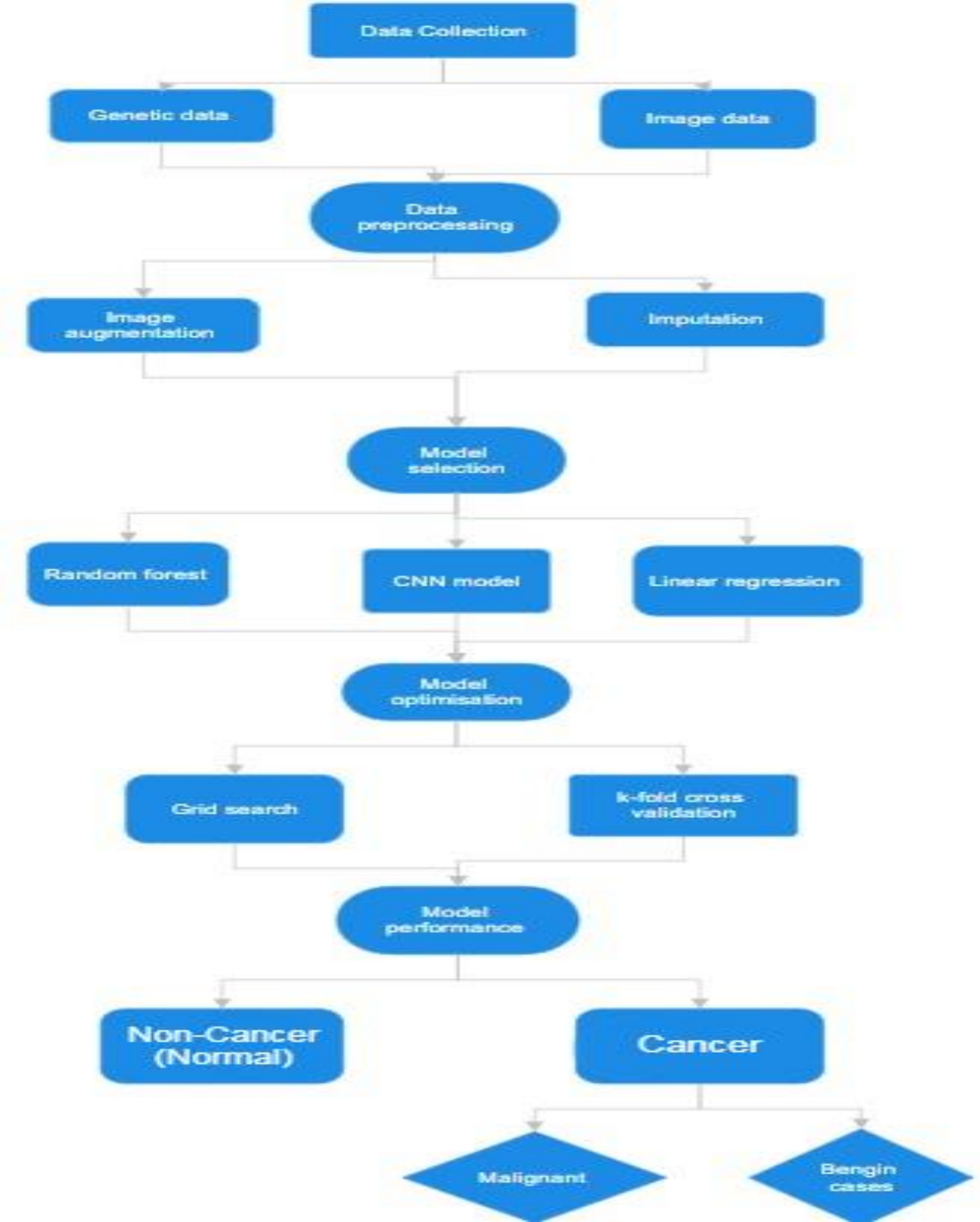


Fig. 1. Workflow of our proposed Model

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI I BENGALURU I CHENNAI I COIMBATORE

METHODOLOGY

DATA COLLECTION

Two datasets were taken from Kaggle that individually serve two different purposes.

- **Survey Lung Cancer dataset:** This dataset consists of various parameters like gender, age, smoking, yellow fingers, lung cancer etc., of many number of patients. This dataset is used in one of the proposed models that uses Linear Regression and Random Forest Classifier to predict whether a person has Lung Cancer or not, in the form of Yes or No.

- **The IQ-OTH/NCCD lung cancer dataset:** This dataset consists of images of the lungs of various patients classified into three cases: Benjin, Malignant and Normal cases. Patients whose lungs are under Benjin and Malignant cases are highly probable of getting affected with Lung cancer whereas the Normal cases are less probable. This dataset is used in the CNN model which is one of the proposed models.

TEAM-12

METHODOLOGY

22BIO211

DATA PREPROCESSING

Random Forest / Linear Regression (Genetic Data)

- Handling Missing Values: Removed or imputed missing entries.
- Encoding Categorical Variables: Used Label Encoding for categorical features like Smoking, Anxiety, Chronic Disease, etc.
- Feature Selection: Applied Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to reduce feature redundancy and improve model efficiency.
- Train-Test Split: Split the dataset into 80% training and 20% testing.
- Feature Scaling: Used StandardScaler to normalize numerical values, ensuring better model performance.

TEAM-12

METHODOLOGY

22BIO211

DATA PREPROCESSING

Convolutional Neural Network (CNN) (Image Data)

- Image Preprocessing:
  - Resized images to 128×128 pixels for uniform input.
  - Applied grayscale conversion (if required).
- Data Augmentation: Used ImageDataGenerator for:
  - Rescaling (1/255) → Normalize pixel values.
  - Rotation, Zoom, Flip, Shift → Improve model generalization.
- Train-Test Split: Used 80% training, 20% validation strategy.
- Normalization: Scaled pixel values between 0 and 1 for stable CNN training.

METHODOLOGY

MODELS USED

1. Random Forest / Linear Regression

Libraries Used: sklearn.ensemble, sklearn.linear_model

- **Random Forest** is an ensemble learning method that builds multiple decision trees and takes the majority vote for classification. It is robust, handles high-dimensional genetic data well, and provides feature importance insights.
- **Linear Regression** was used as a baseline model to understand the relationship between genetic factors and lung disease probability. Though not ideal for classification, it helped in feature influence analysis.
- **Why Used?** Random Forest excels in handling structured data like genetic records, while Linear Regression provides interpretability.

METHODOLOGY

MODELS USED

2. Convolutional Neural Network (CNN)

Libraries Used: tensorflow.keras, tensorflow.keras.layers

- **CNN** is a deep learning model specialized for image classification. It extracts hierarchical features using convolutional layers and classifies lung scans into different categories.
- The model consists of **three convolutional layers with ReLU activation, max pooling layers, and fully connected layers** with softmax activation for multi-class classification.
- **Why Used?** CNN is highly effective in analyzing medical images, detecting patterns that may not be visible through traditional diagnostic methods.

METHODOLOGY

MODEL TRAINING

Random Forest / Linear Regression (Genetic Data)

- Libraries Used: sklearn.ensemble, sklearn.linear_model, sklearn.metrics
- Training Process:
    - **Random Forest Classifier:**
        - Trained on encoded genetic data.
        - Used multiple decision trees to improve classification accuracy.
        - Implemented feature importance analysis to identify key genetic markers.
    - **Linear Regression (Baseline Model):**
        - Used to analyze relationships between genetic features and lung disease probability.
        - Although a regression model, it provided insights into feature influence.

METHODOLOGY

MODEL TRAINING

Random Forest / Linear Regression (Genetic Data)

- Hyperparameter Tuning: Applied GridSearchCV to optimize n_estimators, max_depth, and min_samples_split.
- Evaluation Metrics: Measured Accuracy, Precision, Recall, and F1-Score to assess model performance.

METHODOLOGY

MODEL TRAINING

Convolutional Neural Network (CNN) (Image Data)

- Libraries Used: tensorflow.keras, tensorflow.keras.layers, tensorflow.keras.callbacks
- Training Process:
  - Built a **CNN model** with:
    - **3 convolutional layers (ReLU activation)**
    - **MaxPooling layers** to reduce feature map size
    - Fully connected dense layers with Softmax activation for multi-class classification
  - Trained on preprocessed lung scan images.

TEAM-12

METHODOLOGY

22BIO211

MODEL TRAINING

Convolutional Neural Network (CNN) (Image Data)

- Optimization**:** Used **Adam Optimizer** for efficient learning. Applied **Categorical Crossentropy Loss** since it's a multi-class problem.
- Data Augmentation: Prevented overfitting using rotation, zoom, and flipping techniques.
- Model Checkpointing: Saved the best model using ModelCheckpoint callback.
- Evaluation Metrics: Used Accuracy, Loss, and Confusion Matrix Analysis to evaluate model performance.

RESULTS AND DISCUSSION

Random Forest / Linear Regression (Genetic Data)

- **Random Forest achieved 96.77% accuracy**, effectively classifying lung disease risk based on genetic and clinical data.
- **Feature Importance Analysis** identified smoking history, age, and genetic markers as key predictors.

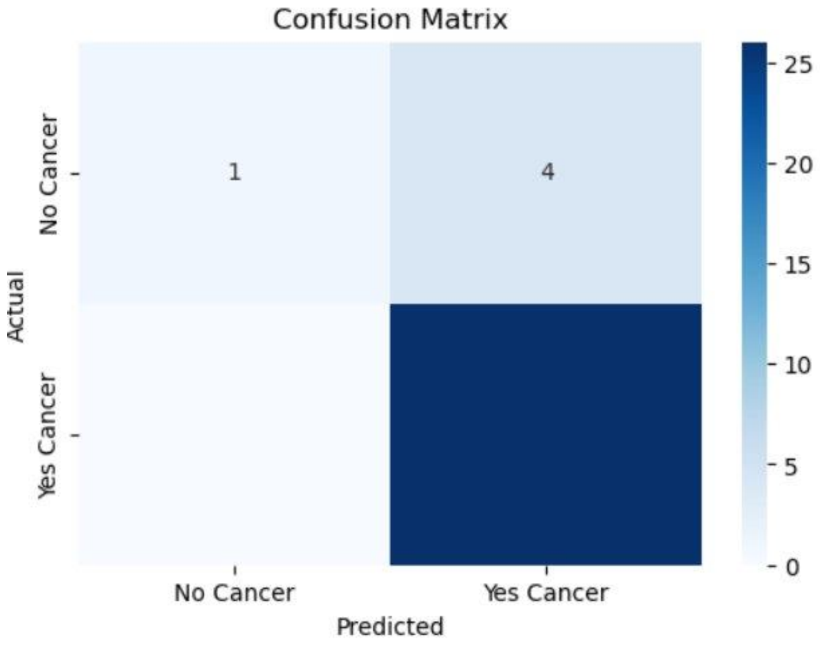
TEAM-12

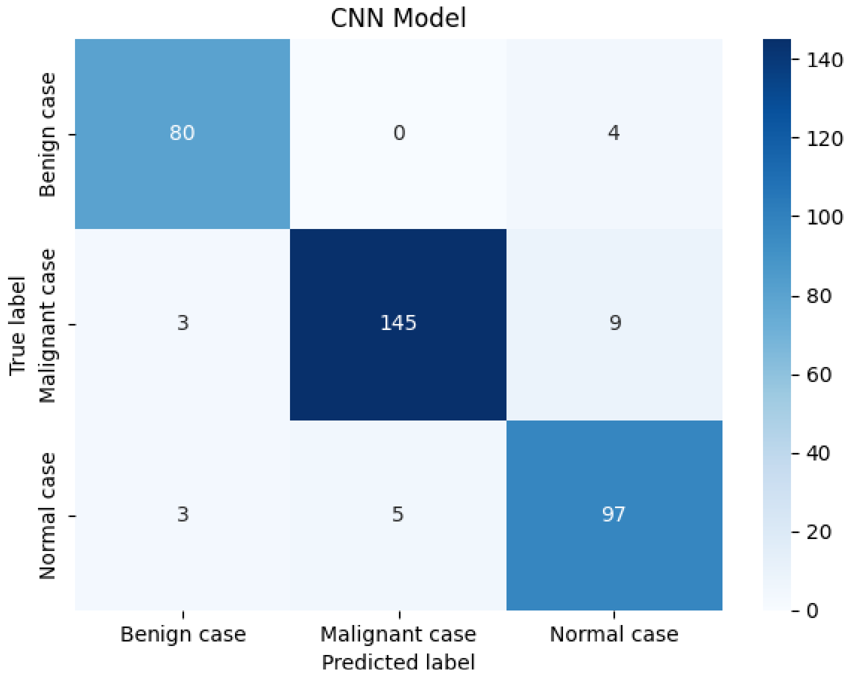
RESULTS AND DISCUSSION

22BIO211

Random Forest / Linear Regression (Genetic Data)

- **Confusion Matrix** showed high sensitivity in detecting lung cancer cases, ensuring fewer false negatives.



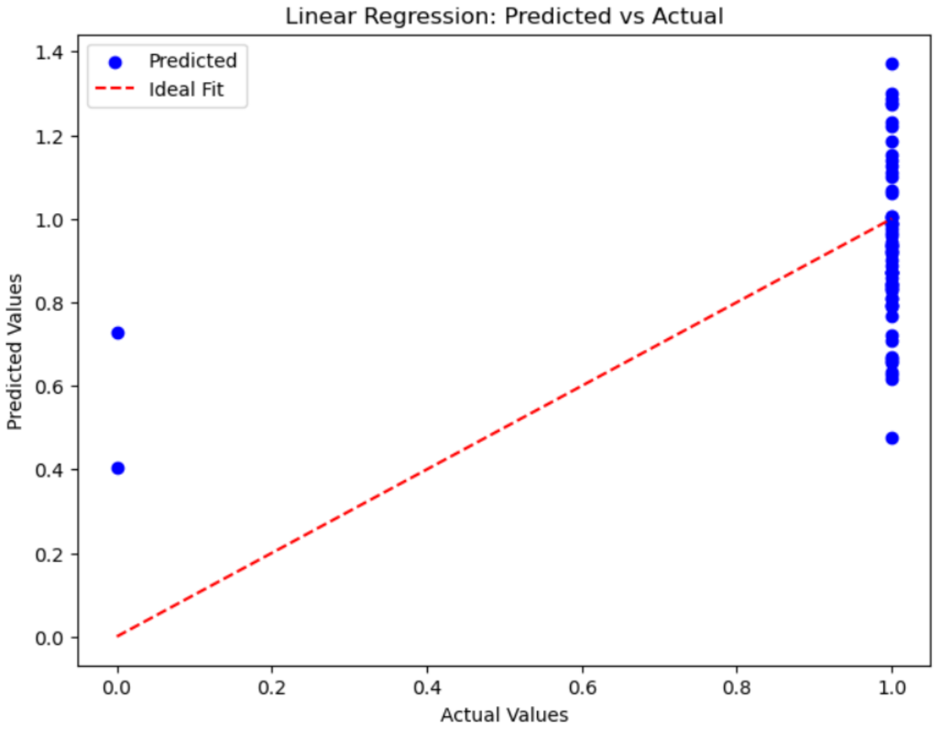


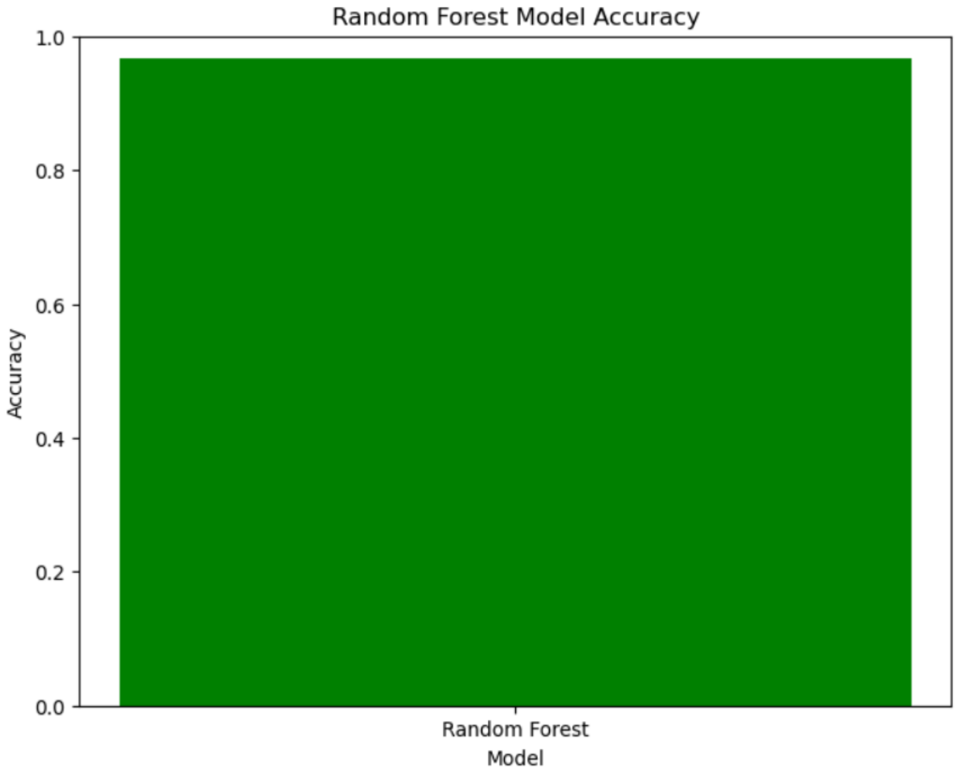
TEAM-12

RESULTS AND DISCUSSION

22BIO211

Random Forest / Linear Regression (Genetic Data)

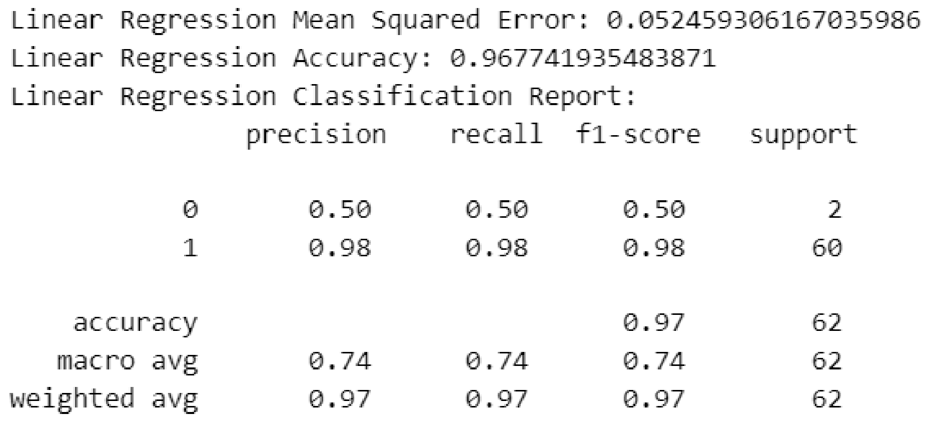




RESULTS AND DISCUSSION

Random Forest / Linear Regression (Genetic Data)

- **Linear Regression** was used as a baseline but was ineffective due to the binary nature of the target variable.
- **Mean Squared Error (MSE) for Linear Regression** confirmed its unsuitability for classification.



```
Linear Regression Mean Squared Error: 0.052459306167035986
Linear Regression Accuracy: 0.967741935483871
Linear Regression Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.50      0.50         2
           1       0.98      0.98      0.98        60

    accuracy                           0.97        62
   macro avg       0.74      0.74      0.74        62
weighted avg       0.97      0.97      0.97        62
```

RESULTS AND DISCUSSION

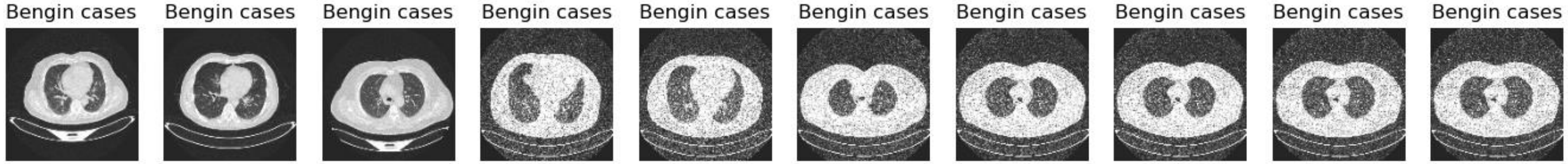
Random Forest / Linear Regression (Genetic Data)

- **Pair Plot Analysis** revealed relationships between numerical features and lung cancer occurrence.
- **Histograms for Numeric Features** provided visual insights into feature distributions.

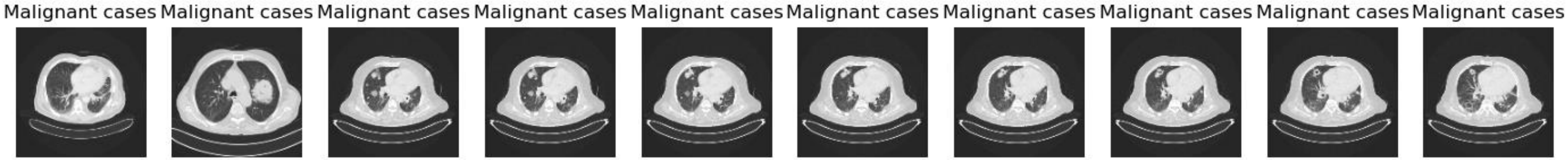
RESULTS AND DISCUSSION

Convolutional Neural Network (CNN)

- **Bengin Case**



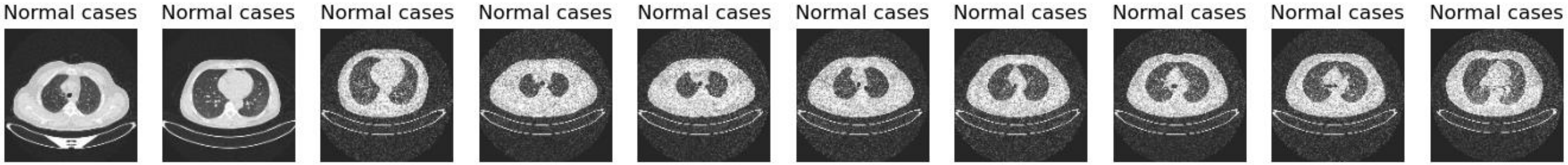
Malignant Case



RESULTS AND DISCUSSION

Convolutional Neural Network (CNN)

Normal Case



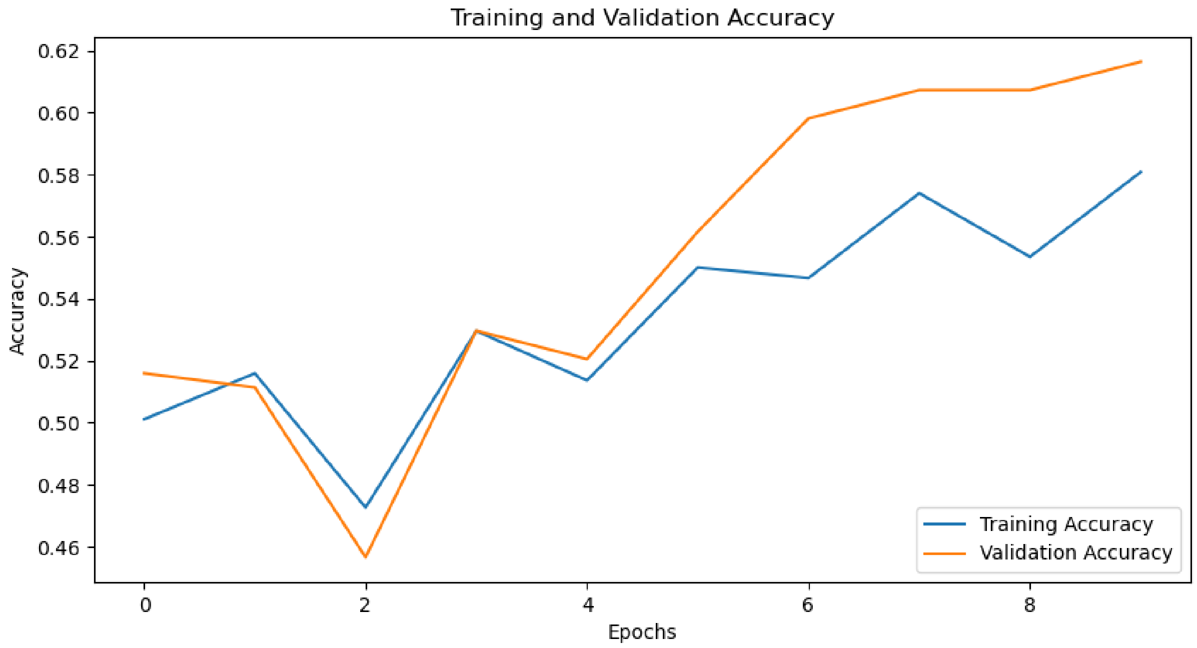
TEAM-12

RESULTS AND DISCUSSION

22BIO211

Convolutional Neural Network (CNN)

- **Training and Validation Accuracy Graph**



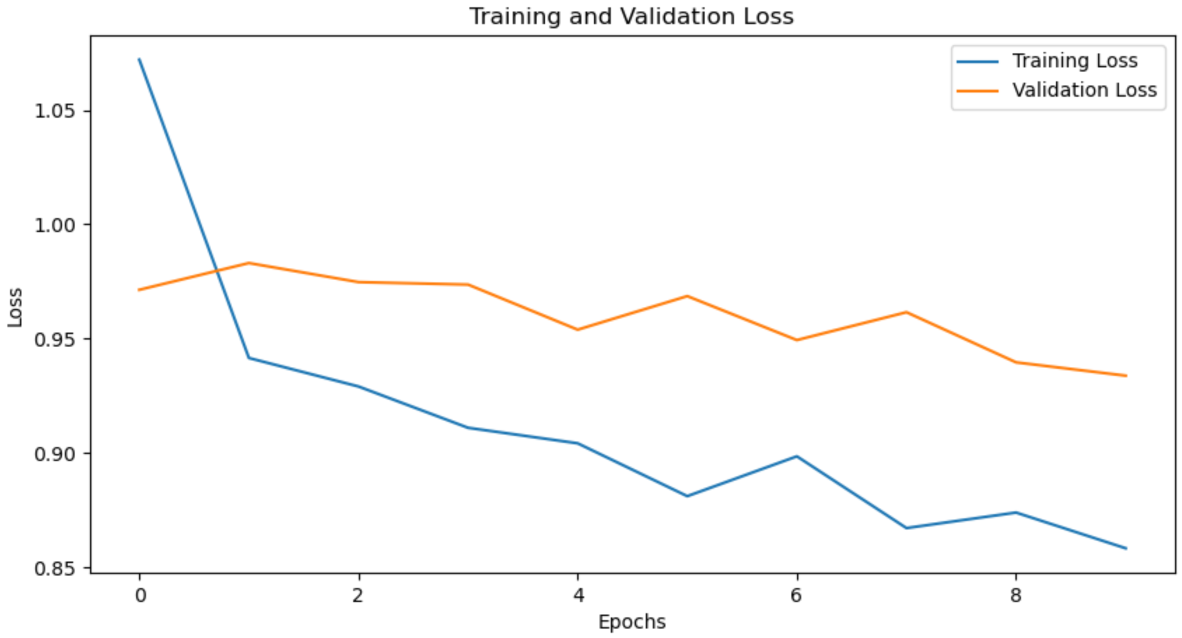
TEAM-12

RESULTS AND DISCUSSION

22BIO211

Convolutional Neural Network (CNN)

- **Training and Validation Loss Graph**



VISUALIZATION

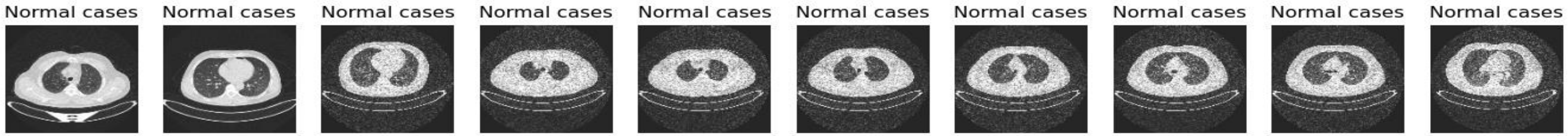


Fig. 3. Represents The Normal Cases

TEAM-12

VISUALIZATION

22BIO211

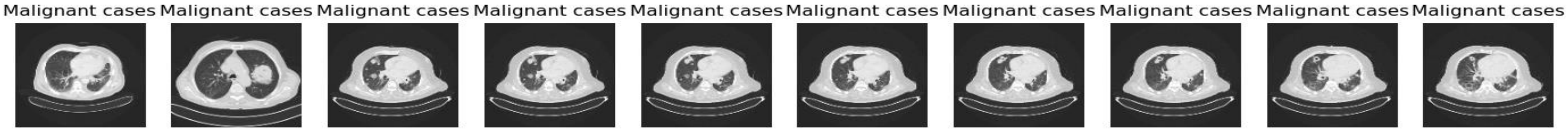


Fig. 4. Represents The Malignant Cases

TEAM-12

VISUALIZATION

22BIO211

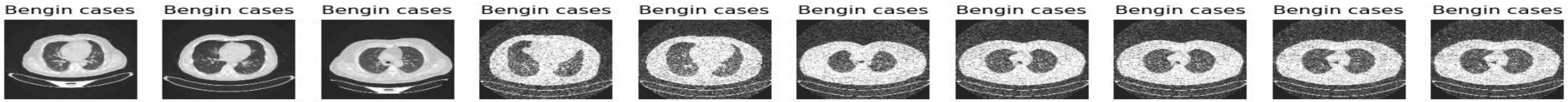


Fig. 5. Represents The Bengin Cases

Literature Review

S.No	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
01	Lung Cancer Prediction Using Machine Learning and Advanced Imaging Techniques **DOI:** 10.21037/tlcr.2018.05.15-	Timor Kadir, Fergus Gleeson **Year:** 2018	Data from publicly available datasets like LIDC-IDRI, which contain annotated lung nodule images. Nodule segmentation was performed using a semi-automated thresholding technique to isolate the region of interest. Radiomics-based texture features, including Haralick and Gabor features, were extracted from both the nodule and its surrounding region.	• Improved classification performance with CNNs • Effective nodule classification using Radiomics • Reduction in false positives using CADx technology	• Significant variability in nodule classification • High false-positive rates remain an issue • Overfitting due to limited training datasets	• Need for large, diverse datasets to improve model generalizability • Lack of integration into clinical workflows • Challenges in applying AI systems across different patient populations

Literature Review

S.No	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
02	Diagnosis of Lung Cancer Based on CT Scans Using CNN DOI: 10.1088/1757-899X/928/2/022035	Hamdalla F. Al-Yasriy, Muayed S. AL-Husieny, Furat Y. Mohsen, Enam A. Khalil, Zainab S. Year-2020	• **Convolutional Neural Network (CNN)** using **AlexNet** architecture. • Applied on lung cancer CT scan images for classification into **normal**, **benign**, or **malignant**. • Training and testing split (70% training, 30% testing).	• Achieved high accuracy of **93.548%**. • High sensitivity (**95.714%**) and specificity (**95%**). • Efficient classification using deep learning for early detection.	• Limited dataset (110 cases). • Model may lack generalization for larger or diverse datasets. • **Research Gap:**	The research gap identified in this study includes the limited dataset size, consisting of only 110 cases, which may restrict the model's ability to generalize effectively to larger and more diverse datasets.

Literature Review

S.No	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
	Prediction and Classification of Lung Cancer Using Machine Learning Techniques DOI: 10.1088/1757-899X/1099/1/012059	Pragya Chaturvedi, Anuj Jhamb, Meet Vanani, Varsha Nemade Published in MDPI - 2024 Year-2021	• Various **machine learning techniques** for lung cancer classification. • **Image Preprocessing:** Gaussian filtering, median filtering, adaptive bilateral filtering. • **Segmentation Methods:** Watershed transform, Sobel edge detection, K-Means clustering. • **Classification Algorithms:** CNN, SVM, KNN,	• **Comprehensive comparison** of multiple segmentation and classification techniques. • **High classification accuracy** with CNN (up to 97.2%). • Effective **feature extraction** using hybrid approaches.	• **Limited dataset** used, which may impact generalizability. • Certain **segmentation techniques** have high computational costs. • Lacks **real-time clinical validation.**	• The study primarily focuses on **technical comparison** without **real-world deployment.** • **Lack of large-scale validation** across diverse datasets. • Requires **integration with CAD systems** for improved automation in medical diagnosis.

Literature Review

S.No	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
	Deep Learning Ensemble 2D CNN Approach Towards the Detection of Lung Cancer DOI: 10.1038/s41598-023-29656-z	Asghar Ali Shah, Hafiz Abid Mahmood Malik, AbdulHafeez Muhammad, Abdullah Alourani, Zaeem Arif Butt Year-2023	• **Deep Ensemble 2D CNN** consisting of three different CNN models. • **LUNA 16 Grand Challenge** dataset used for training and testing. • Each CNN uses different layers, kernels, and pooling techniques for improved accuracy. • Combined prediction using an ensemble approach.	• Achieved **95% accuracy**, higher than baseline methods. • Reduced false positives with ensemble learning. • Effective feature extraction using deep CNNs.	• Requires **high computational resources**. • Limited to 2D CNN, lacking spatial information from 3D images.	• The study focused only on **2D CNN**; exploring **3D CNN** could enhance accuracy. • Lack of **real-world clinical validation**. • Further testing on larger and more diverse datasets is necessary for generalization.

Literature Review

S.No	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
	Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework DOI: 10.31661/jbpe.v0i0.1027	Shayesteh S. P., Shiri I., Karami A. H., Hashemian R., Kooranifar S., Ghaznavi H., Shakeri-Zadeh A. Year-2020	• **Logistic Regression-based Models** for survival time prediction. • **Radiomics Framework** for feature extraction from CT images. • **Mutual Information-based Feature Selection** to choose relevant features. • **Six Logistic Regression Models** including Dual Coordinate Descent (DCD-LR).	• Provides a **non-invasive** method for predicting survival time. • Effective **feature selection** using mutual information. • DCD-LR model demonstrated the **best accuracy and F1 score**.	• **Limited dataset** of 59 patients, affecting model generalizability. • Potential for **overfitting** with a small number of selected features.	• **Lack of external validation** with larger datasets. • **Limited comparison** with other non-logistic regression models. • Future research could integrate genomic and radiomic data for improved predictions.

Future Works

- Integration of Genetic and Imaging Data: Combining structured genetic data with medical imaging can improve lung disease prediction accuracy.
- Enhancing Model Interpretability: Implementing Explainable AI (XAI) techniques like Grad-CAM for CNN and SHAP values for Random Forest can improve trust in model decisions.
- Addressing Data Imbalance: Applying advanced resampling techniques like SMOTE for genetic data and GANs for medical images can enhance model performance.
- Optimizing CNN Architectures: Experimenting with deeper models like ResNet or EfficientNet can improve classification accuracy for Malignant, Benign, and Normal cases.
- Real-World Clinical Validation: Testing the model on larger, diverse datasets and collaborating with healthcare institutions can validate its effectiveness for practical applications.
- Edge Deployment for Faster Diagnosis: Implementing the trained model on embedded devices for real-time lung disease detection in clinical settings

CONCLUSION

This study demonstrated that **Random Forest excels in genetic data analysis**, while **CNN effectively classifies lung scans into Malignant, Benign, and Normal cases**. The integration of **AI-driven models** enhances **early lung disease detection, risk assessment, and diagnostic accuracy**. Despite challenges like **data imbalance and model interpretability**, future improvements in **multi-modal AI and real-world clinical validation** can further optimize disease prediction and patient care