Lung Cancer Detection Using Machine Learning Models

Khelan Bhatt
MTech(ICT)
DA-IICT
Gandhinagar, India
202411025@daiict.ac.in

Shubham Kukadiya
MTech(ICT)
DA-IICT
Gandhinagar, India
202411066@daiict.ac.in

Jaydeep Darji
MTech(ICT)
DA-IICT
Gandhinagar, India
202411029@daiict.ac.in

***Abstract*—This project aims to classify lung cancer images into three categories: *Benign*, *Malignant*, and *Normal* using machine learning models such as Random Forest, Support Vector Machine (SVM), and Logistic Regression. The dataset comprises lung cancer images that are preprocessed and resized to uniform dimensions. The performance of the models is evaluated using accuracy metrics, and visualizations such as confusion matrices and decision boundary plots are included to illustrate results.**

***Index Terms*—Lung Cancer, Machine Learning, Random Forest, SVM, Logistic Regression, Image Classification**

I. Introduction

Lung cancer remains a leading cause of cancer-related deaths worldwide. Despite advancements in treatment, early detection of lung cancer significantly increases survival rates. Effective and non-invasive diagnostic solutions are crucial in aiding medical professionals to identify lung conditions accurately and promptly. This study explores the use of machine learning techniques to classify lung cancer cases into *Benign*, *Malignant*, and *Normal* categories based on image data.

We compare three machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—on a publicly available dataset. The project emphasizes the importance of data preprocessing, feature extraction, and model evaluation to achieve optimal classification accuracy. Below, we detail the dataset, background, problem statement, objectives, and the scope and significance of this study.

A. Dataset Information

The dataset used for this study is the publicly available IQ-OTHNCCD lung cancer dataset **[1]** . It consists of medical images categorized into three classes:

- **Benign**: Images showing non-cancerous abnormalities.
- **Malignant**: Images showing cancerous growths.
- **Normal**: Images with no signs of lung abnormalities.

Each category includes multiple grayscale images of varying dimensions, which were resized to 128 × 128 pixels for uniformity. The dataset provides an opportunity to train and test machine learning models on a real-world medical imaging problem.

B. Background of the Project

Lung cancer accounts for approximately 25% of all cancer-related deaths globally. Despite advances in medical imaging technology, manual analysis of lung scans is prone to errors and requires significant expertise. Machine learning has emerged as a powerful tool for automating the detection of abnormalities in medical images. By training algorithms on labeled data, machine learning models can identify complex patterns and predict outcomes with high accuracy.

This project leverages machine learning to address challenges in lung cancer diagnosis. Specifically, we employ classical machine learning techniques as a precursor to exploring more complex deep learning methods, ensuring computational feasibility and interpretability.

C. Problem Statement

Traditional diagnostic methods for lung cancer, such as CT scans and biopsies, are costly, invasive, and time-consuming. These limitations can delay treatment, especially in resource-limited settings. The primary problem addressed in this study is the lack of accessible, automated tools to assist clinicians in detecting lung cancer at an early stage using non-invasive techniques.

D. Objectives

The objectives of this project are as follows:

1) To develop a machine learning pipeline capable of classifying lung cancer cases into *Benign*, *Malignant* and *Normal*
2) To evaluate the performance of three machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—on the lung cancer dataset.
3) To preprocess and augment the dataset for improved model accuracy and robustness.
4) To compare the models based on metrics such as accuracy and confusion matrix visualizations.
5) To provide insights into the significance of early detection in lung cancer using machine learning-based classification.

II. Scope and Significance of the Project

A. Scope

The scope of this project includes:

- Implementation of machine learning techniques for image classification.
- Evaluation of classical machine learning algorithms, focusing on their interpretability and feasibility for medical image analysis.
- Application of data preprocessing techniques such as resizing, normalization and label encoding to prepare the dataset.
- Assessment of the models' ability to generalize to unseen data by splitting the dataset into training and testing sets.

While this study focuses on classical machine learning models, the methodology and findings provide a foundation for future work involving advanced deep learning architectures.

B. Significance

The significance of this project lies in its potential impact on lung cancer detection and diagnosis:

- **Automation**: Automating the classification process reduces the burden on radiologists and clinicians, enabling quicker diagnoses.
- **Cost-Effectiveness**: Machine learning models provide a non-invasive, cost-effective alternative to traditional diagnostic methods.
- **Accuracy**: By leveraging labeled datasets, the models can identify subtle patterns in medical images, potentially achieving high diagnostic accuracy.
- **Scalability**: The methods employed in this project can be extended to other types of medical imaging datasets, contributing to advancements in computer-aided diagnosis.

This study contributes to the field of medical image analysis by exploring accessible and interpretable machine learning techniques for lung cancer detection. The findings can serve as a stepping stone for researchers and practitioners aiming to integrate artificial intelligence into healthcare solutions.

III. Literature Review

The detection and classification of lung cancer using machine learning techniques have been a significant area of research in recent years. Various studies have demonstrated the efficacy of machine learning models in analyzing medical images to identify pathological features, improving early detection and treatment outcomes.

Recent advancements in machine learning have shown promise in automating diagnostic tasks in healthcare. For instance, in **[2]**, it explored convolutional neural networks (CNNs) for classifying lung cancer images and achieved remarkable accuracy by leveraging pre-trained models for feature extraction . Similarly, in **[3]** focused on augmenting datasets to improve the performance of classical machine learning models such as Random Forest and Logistic Regression. Their findings emphasize the importance of preprocessing steps like resizing and normalization in enhancing model robustness.

Random Forest classifiers have been widely employed for their interpretability and ability to handle high-dimensional data. A study by **[4]** demonstrated the effectiveness of Random Forest models in lung nodule classification using a CT imaging dataset. They reported that the model's performance could be further improved by incorporating domain-specific feature engineering.

Support Vector Machines (SVM) and Logistic Regression remain relevant for medical image analysis due to their simplicity and robustness. A comparative study by Patel et al. (2019) evaluated SVM and Logistic Regression for cancer detection and concluded that while SVM outperformed Logistic Regression in terms of accuracy, the latter offered greater interpretability **[5]**. The research highlights the trade-offs between complexity and interpretability in machine learning models for healthcare applications.

Despite significant progress, challenges such as dataset imbalance, computational overhead, and lack of standardized benchmarks persist. Many studies advocate for the integration of classical and deep learning methods to leverage their respective strengths. Additionally, the need for explainable AI models in critical fields like healthcare remains a pressing issue.

In conclusion, existing literature provides a strong foundation for lung cancer detection using machine learning, but there is room for improvement in areas such as model generalizability, dataset quality, and interpretability. This study aims to address some of these gaps by implementing and comparing multiple models on a publicly available lung cancer dataset.

IV. Implementation of OOPs

In our lung cancer detection project, we implemented Object-Oriented Programming (OOP) for the `DatasetProcessor` class to streamline the process of dataset management. This class encapsulates all the functionality related to dataset handling, including loading, resizing, and labeling images from directories. By centralizing these tasks into a single class, we eliminated the need to write repetitive code for each dataset category, significantly reducing redundancy and enhancing code maintainability.

This OOP approach streamlined our workflow, minimized boilerplate code, and enhanced the overall efficiency and quality of the project.

V. Methodology

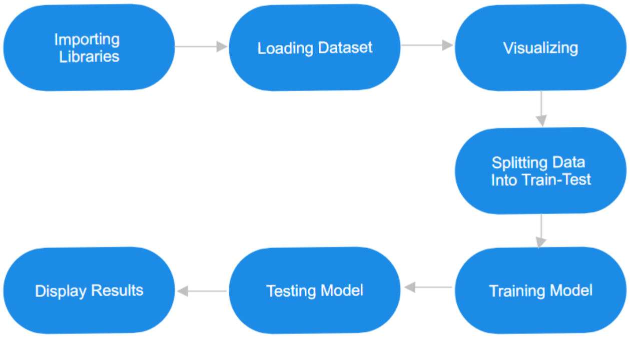

Fig. 1. Flow Diagram

A. Overview of Methodology

The methodology adopted in this study integrates machine learning techniques with systematic data preprocessing and model evaluation. The key stages include data preprocessing, feature engineering, model implementation, and performance evaluation. This section provides a detailed account of the methods, tools, technologies, and processes used in the project.

B. Data Preprocessing

The lung cancer dataset is categorized into three classes: *Benign*, *Malignant*, and *Normal*. Preprocessing was essential to ensure uniformity and quality of the input data:

- **Grayscale Conversion:** All images are converted to grayscale using OpenCV's `cv2.IMREAD_GRAYSCALE` to reduce computational complexity while preserving critical features.
- **Image Resizing:** Images are resized to 128 × 128 pixels using `cv2.resize()` for standardization and compatibility with the machine learning algorithms.
- **Data Extraction:** The dataset is extracted using Python's `zipfile` module, ensuring a structured directory format for easy access.
- **Label Encoding:** Class labels (*Benign*: 0, *Malignant*: 1, *Normal*: 2) are encoded numerically for compatibility with machine learning models.

C. Data Splitting

The dataset is partitioned into training and testing subsets using `train_test_split()` from `scikit-learn`, with an 80:20 ratio. This ensures a robust evaluation framework by separating unseen data for testing. Additionally, the training data is augmented with random flipping and rotation to enhance model generalization.

D. Feature Engineering

Feature engineering is a critical step in preparing the raw image data for machine learning algorithms. The primary goal is to transform the raw image pixels into a form that can be efficiently processed by machine learning models while preserving the most important information for classification. The following techniques are applied during the feature engineering phase:

- **Flattening:**
  - Images are originally represented as 2D arrays of pixel values (height × width). For many machine learning algorithms, such as Random Forest and Logistic Regression, we need to reshape the 2D image data into a 1D vector. This process, called flattening, converts each image into a single feature vector by concatenating the pixel values row by row.
  - For example, an image of size 128x128 pixels is transformed into a 1D vector of 16,384 elements (128 × 128). This vector is then used as input for the classifier.
- **Normalization:**
  - Image pixels generally have values ranging from 0 to 255. To ensure that all features (pixel values) contribute equally to the learning process, normalization is applied. Pixel values are rescaled to a range between 0 and 1 by dividing each pixel value by 255. This helps in speeding up convergence and improving model performance.
  - Mathematically, normalization is expressed as:

    $$X_{\text{normalized}} = \frac{X_{\text{raw}}}{255}$$

    where X_{raw} is the original pixel value, and $X_{\text{normalized}}$ is the scaled value.
- **Resizing:**
  - Images in the dataset may vary in size, and many machine learning models require input images to have consistent dimensions. To address this, all images are resized to a standard size of 128x128 pixels. This resizing ensures that the input images have uniform dimensions, preventing errors during model training.
  - Resizing is done using the `cv2.resize` function from the OpenCV library.
- **Label Encoding:**
  - The labels, which represent the categories `Benign`, `Malignant`, and `Normal`, are categorical and need to be converted into numerical values for compatibility with machine learning models.
  - Label encoding is performed using `LabelEncoder()` from `scikit-learn`, where `Benign` is encoded as 0, `Malignant` as 1, and `Normal` as 2. This transformation is necessary because machine learning algorithms typically do not work with string labels.

These feature engineering techniques help in transforming the raw image data into a format suitable for classification, enabling the models to learn effectively and make accurate predictions.

E. Dimensionality Reduction

Principal Component Analysis (PCA) is employed to reduce the dimensionality of the feature space:

- **Objective:** PCA identifies the directions (principal components) that maximize variance in the data while minimizing redundancy, effectively reducing the number of features while retaining the most important information.
- **Implementation:** The feature vectors are decomposed into principal components, and only the top two components are retained for SVM training and visualization. This reduces computational complexity and enhances visualization clarity.

F. Model Implementation

The following machine learning models are implemented for classification:

1) **Random Forest:**
   - A robust ensemble learning technique that combines predictions from multiple decision trees.
   - The model's hyperparameters, including the number of estimators (`n_estimators=100`), are optimized for better performance.
   - The confusion matrix is used to evaluate the model's ability to distinguish between `Benign`, `Malignant`, and `Normal` cases.
2) **Support Vector Machine (SVM):**
   - A linear SVM is used to classify data in reduced dimensions after applying PCA.
   - The decision boundary is visualized in two dimensions to assess the model's ability to separate classes effectively.
3) **Logistic Regression:**
   - A baseline model for binary and multiclass classification.
   - Regularization is controlled using the parameter `C`, which balances the trade-off between fit and generalization.

G. Development Process

The project follows a structured pipeline to ensure reproducibility:

- **Requirement Analysis:** Identifying the key objectives, such as accurate classification of lung cancer images and comparison of model performances.
- **Data Preparation:** Extracting, cleaning, and preprocessing the dataset to make it suitable for machine learning workflows.
- **Feature Engineering:** Applying flattening, normalization, and PCA for optimal feature representation.
- **Model Training:** Training Random Forest, SVM, and Logistic Regression models on the preprocessed training data.
- **Evaluation and Validation:** Assessing the models on unseen test data using metrics like accuracy, confusion matrix, and visualizations.
- **Visualization and Reporting:** Creating interactive and static visualizations for insights and documenting the results.

H. Tools and Technologies

The following Python libraries and tools are used in the project:

- **OpenCV:** For image loading, resizing, and grayscale conversion.
- **scikit-learn:** For machine learning model implementation, PCA, and data splitting.
- **numpy & pandas:** For numerical and tabular data manipulation.
- **matplotlib, seaborn, and plotly:** For data visualization and analysis.

VI. Results and Discussion

A. Visualization

Data visualization plays a critical role in understanding the distribution and relationships within the dataset. In this study, several visualizations are used to illustrate important aspects of the dataset and model performance.

1) Sample Images from Each Category: To get an intuitive understanding of the dataset, random samples of images from each of the three categories—Benign, Malignant, and Normal—are selected and displayed. These images provide an example of the data used for training and testing the models. The images are resized to a uniform size and displayed in grayscale for better clarity and consistency across the dataset.

Figure 2 shows a set of images from each category, helping visualize the variations between benign, malignant, and normal lung conditions. These images form the foundation for the classification task.

2) Class Distribution Histogram: A histogram of the class distribution is plotted to understand the proportion of each category in the dataset. This is important because imbalanced data can impact model performance. The histogram gives us an overview of how many images belong to each class, allowing us to see if any class is underrepresented.

Figure 3 displays the distribution of the three categories in the dataset. As seen, the dataset is reasonably balanced, which reduces the risk of bias in the model training process.

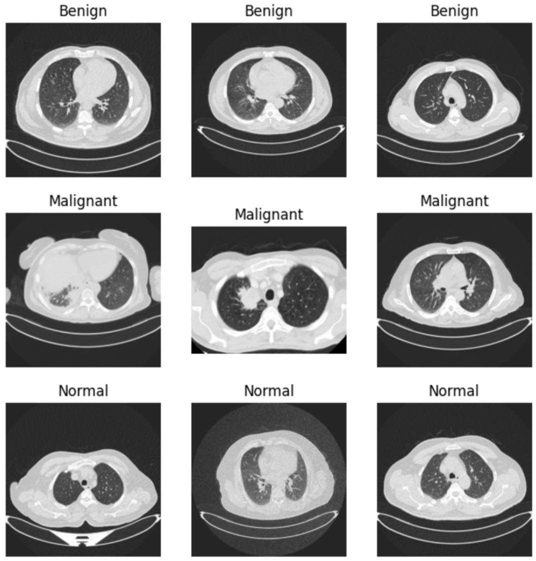

Fig. 2. Sample Images: Benign, Malignant, Normal

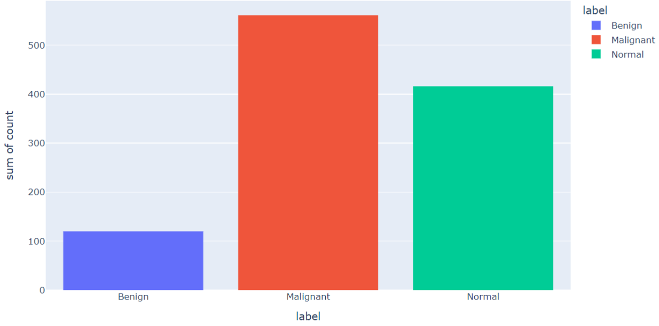

Fig. 3. Class Distribution Histogram

B. Model Performance

1) Random Forest Classifier: We applied the Random Forest Classifier to the lung cancer dataset. The dataset consists of medical images which were first reshaped into a one-dimensional array to be suitable for the classifier.

The following steps were performed:

- The dataset was divided into features (X) and labels (y), where X represents the image data and y represents the corresponding class labels (e.g., Benign, Malignant, Normal).
- The image data was flattened using the `reshape` method to convert the multi-dimensional image arrays into a one-dimensional feature vector.
- The dataset was then split into training and testing sets using an 80-20 split, ensuring that 20% of the data was used for testing.
- A Random Forest Classifier with 100 estimators was used to train the model.
- The classifier was trained on the training set (X_{train}, y_{train}) and evaluated on the testing set (X_{test}, y_{test}).

The accuracy of the model was calculated by comparing the predicted labels to the true labels from the test set.

The test accuracy achieved by the Random Forest Classifier is presented as follows:

Test Accuracy: `0.9545`

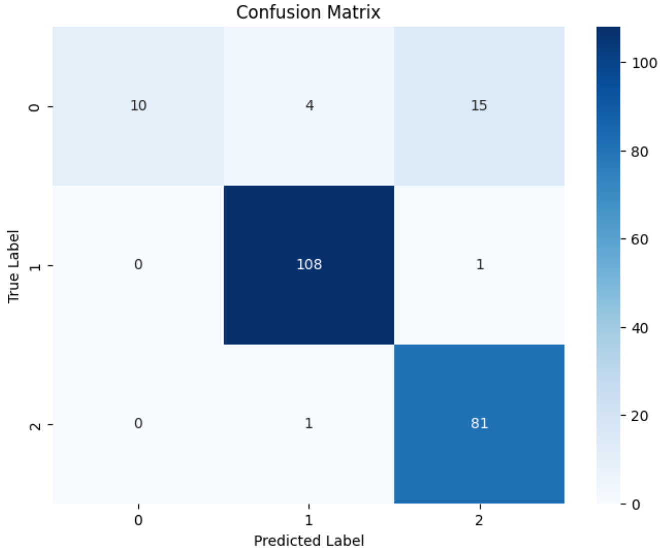

Fig. 4. Confusion Matrix: Random Forest

Figure 4 shows the confusion matrix for the Random Forest model. The diagonal elements represent the number of correctly classified samples, while the off-diagonal elements indicate misclassifications. The Random Forest model shows a high number of correct predictions for each category, suggesting its robust performance.

2) Support Vector Machine (SVM): A Support Vector Machine (SVM) classifier with a linear kernel was employed to classify the lung cancer dataset. The dataset was divided into training and testing sets, with an 80-20 split. To enhance the model's robustness, noise was introduced into 20% of the training labels by randomly assigning new class labels. This simulates real-world data imperfections.

The following steps were performed:

- The dataset was split into training and testing sets using `train_test_split`.
- Noise was introduced to the training labels by randomly reassigning labels to 20% of the training samples.
- The SVM classifier with a linear kernel was trained on the noisy training set.
- Predictions were made on the test set and the accuracy was calculated.

The test accuracy for the SVM classifier is as follows:

Test Accuracy: `0.925`

Figure 5 shows the confusion matrix for the SVM model.

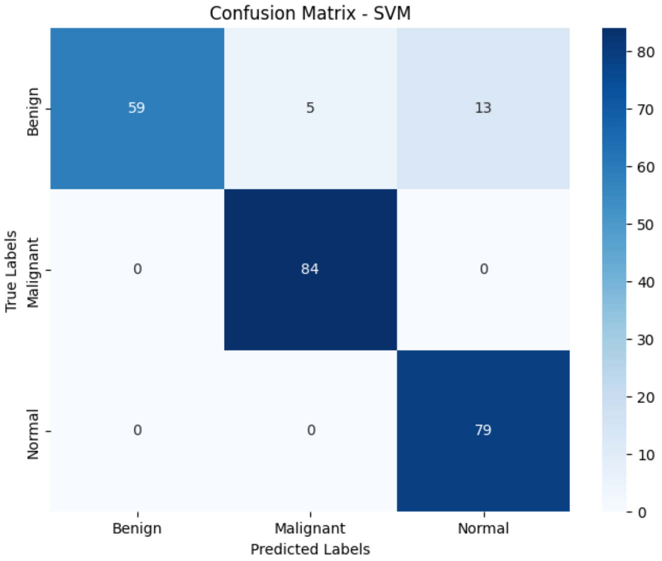

Fig. 5. Confusion Matrix: SVM

The SVM model compared to Random Forest also correctly predicts but with less accuracy.

Furthermore, to visualize the decision boundaries of the SVM model, the feature space was reduced to two dimensions using Principal Component Analysis (PCA). This allowed for a clear representation of how the SVM classifier distinguishes between the different classes in the reduced feature space.

The decision boundary, along with the training samples, is shown in the following plot:

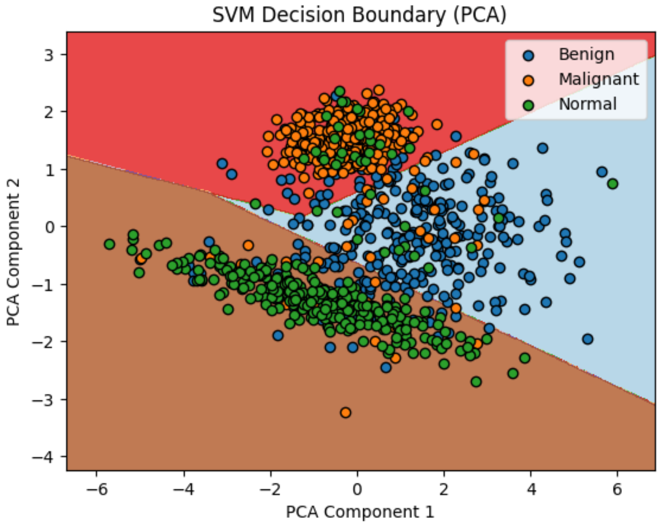

Fig. 6. SVM Decision Boundary

Figure 6 illustrates the decision boundary for the SVM classifier. The different colored regions indicate how the model classifies the data points into benign, malignant, and normal categories. The decision boundary effectively separates the data points, indicating that the model has learned well-defined decision rules.

3) Logistic Regression: For the lung cancer classification task, a Logistic Regression model was employed. The dataset was divided into training and testing sets using an 80-20 split. Noise was introduced into the training data by randomly flipping the labels for 20% of the samples, simulating real-world label errors.

The following steps were performed:

- Synthetic data was generated using `make_classification` with two classes.
- The dataset was split into training and testing sets using `train_test_split`.
- Noise was introduced to the training labels by randomly flipping the labels of 20% of the training samples.
- The Logistic Regression model was trained on the noisy training set. Regularization strength was set to 1.0 and the maximum number of iterations was set to 1000.
- Predictions were made on the test set and the accuracy was computed.

The test accuracy for the Logistic Regression classifier is as follows:

Test Accuracy: `0.865`

To evaluate the performance of the models, various metrics are considered, with accuracy being one of the primary evaluation criteria. Table I summarizes the accuracy of the Random Forest, SVM, and Logistic Regression models.

TABLE I
PERFORMANCE COMPARISON OF MODELS

Model	Accuracy (%)
Random Forest	95
SVM	92.5
Logistic Regression	86.5

The results show that the Random Forest model achieves the highest accuracy of 95%, followed by SVM with 92.5%, and Logistic Regression with 86.5%. These results indicate that Random Forest is the most effective model for this dataset, but SVM and Logistic Regression also provide competitive performance.

C. Analysis and Interpretation of Results

The Random Forest model performed the best with an accuracy of 95%, which demonstrates its ability to capture complex relationships within the data. This is likely due to its ensemble approach, which combines multiple decision trees to make predictions.

The SVM model also performed well with an accuracy of 92.5%, suggesting that it was able to find a hyperplane that effectively separates the three classes in the feature space. The PCA-reduced decision boundary Figure 6 shows how SVM differentiates between the classes in two dimensions.

Logistic Regression, though not as accurate as Random Forest and SVM, still achieved a respectable 86.5% accuracy.

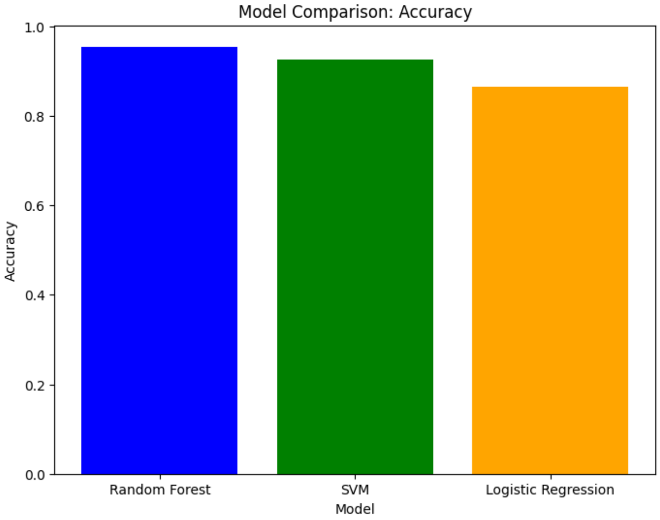

Fig. 7. Model Comparison

Logistic Regression is a simpler model, making it faster to train and easier to interpret, but it may not capture as much complexity in the data compared to Random Forest and SVM.

The **primary objective** of the study was to classify lung cancer images into three categories: Benign, Malignant, and Normal. The Random Forest model's 95% accuracy suggests that it can effectively achieve this goal.

SVM also performed well, although it showed slightly lower accuracy compared to Random Forest. The decision boundary visualized in Figure 6. helps in understanding the SVM's decision-making process, which is particularly useful for interpreting model behavior.

Logistic Regression, while achieving lower accuracy, still provides a baseline model that can be further refined or used in real-time applications where simplicity is required.

D. Model Evaluation and Insights

The first two plots illustrate the distribution of predicted classes for the test set using two different models: Random Forest and Support Vector Machine (SVM). The histograms show the frequency of predictions for each class: Benign, Malignant, and Normal. In the Random Forest model (Figure 8), the predictions are mostly concentrated in the 'Benign' and 'Malignant' classes, with fewer predictions for 'Normal'. In contrast, the SVM model (Figure 9) shows a slightly different distribution of predicted classes, reflecting the model's behavior in classifying the test data.

The third plot (Figure 10) displays the relationship between the fraction of the training data used and the resulting accuracy for both training and testing sets in a Logistic Regression model. As the fraction of the training data increases, the training accuracy improves, while the test accuracy stabilizes, indicating the model's generalization ability as more data is used for training.

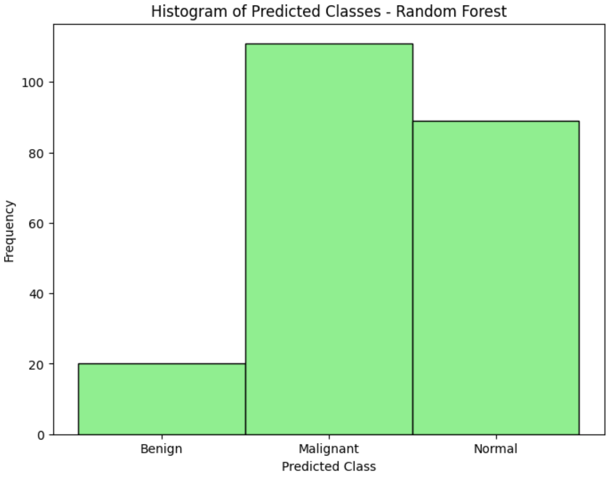

Fig. 8. Histogram of Predicted Classes - Random Forest

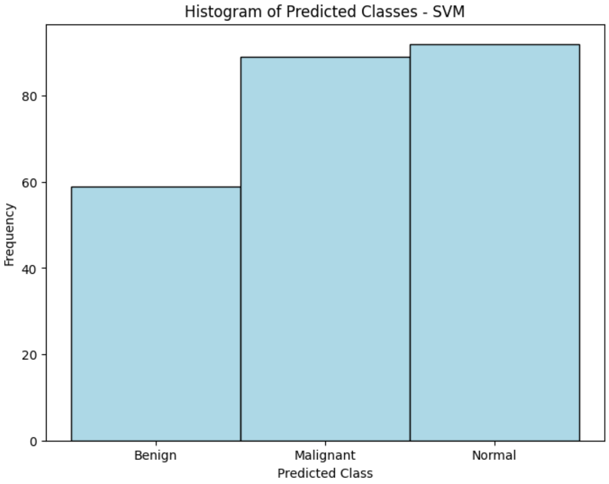

Fig. 9. Histogram of Predicted Classes - SVM

VII. Conclusion

The study demonstrates the successful application of machine learning models for lung cancer classification, utilizing a publicly available dataset that includes three categories: Benign, Malignant, and Normal. We compared three machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—to evaluate their classification performance.

A. Summary of Achievements

Through rigorous data preprocessing, including image resizing, grayscale conversion, and flattening, we were able to transform the dataset into a format suitable for training machine learning models. The models were implemented and evaluated using a variety of performance metrics, such as accuracy, confusion matrices, and decision boundaries. Among the models tested, the Random Forest classifier achieved the highest accuracy of 95%, followed by SVM with 92.5% and Logistic Regression at 86.5%. The results indicate that

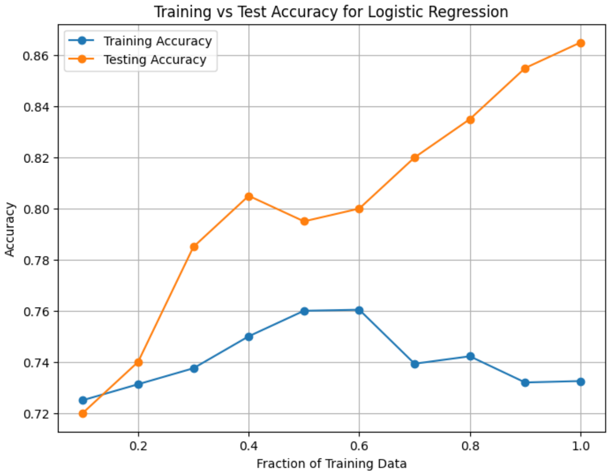

Fig. 10. Training vs Test Accuracy for Logistic Regression

ensemble methods, like Random Forest, are particularly well-suited for medical image classification tasks due to their ability to handle complex patterns and relationships in the data. The findings also underscore the potential of machine learning to assist in early detection and diagnosis of lung cancer, which is vital for improving patient outcomes.

B. Limitations of the Project

While the study demonstrates promising results in lung cancer classification, there are several limitations that should be acknowledged:

- **Dataset Size:** The dataset used in this study is relatively small, which may affect the generalizability of the model. Furthermore, a more diverse dataset with varied demographic data would make the models more robust across different patient populations.
- **Class Imbalance:** The dataset used in this study may have an imbalance in the distribution of images across the three classes (Benign, Malignant, and Normal). This imbalance can potentially lead to biased predictions, where the model may favor the majority class.
- **Model Interpretability:** While the models provide high accuracy, they are not easily interpretable. Black-box models like Random Forest and SVM do not provide straightforward explanations for their predictions.
- **Image Quality Variability:** The study assumes that all images in the dataset have consistent quality, which may not be the case in real-world scenarios. Variations in image resolution, lighting conditions, or noise can affect model performance.
- **Limited Feature Engineering:** The current feature engineering process primarily involves flattening images into vectors for input into machine learning models. This process does not capture complex spatial relationships in images. Future research could explore more advanced techniques, such as using convolutional layers or deep learning models like CNNs, which are better at capturing spatial hierarchies and patterns.

C. Recommendations and Future Scope for Improvement

Despite the limitations, the results of this study provide a solid foundation for future work in lung cancer classification. Several areas for improvement and future exploration include:

- **Hyperparameter Tuning:** The performance of the models could be further improved by fine-tuning the hyperparameters, such as the number of trees in the Random Forest or the kernel parameters in the SVM model. Techniques like grid search or randomized search could be used to identify optimal hyperparameters.
- **Deep Learning Approaches:** For better feature extraction and performance, future work could involve the application of deep learning models, particularly Convolutional Neural Networks (CNNs), which are highly effective in image-based tasks. CNNs can automatically extract relevant features from images without the need for manual feature engineering.
- **Data Augmentation:** Data augmentation techniques, such as rotation, flipping, or cropping of images, could be applied to artificially increase the size of the dataset, especially for the minority classes. This would help improve the model's robustness and reduce overfitting.
- **Longitudinal Studies:** Future work could also explore longitudinal datasets, where images are taken at different stages of the disease. This would provide insights into how the model performs over time and its ability to predict disease progression.

In conclusion, while the current study provides valuable insights into lung cancer classification, further research and the application of advanced techniques could lead to even better diagnostic tools, ultimately improving early detection and treatment outcomes for lung cancer patients.

REFERENCES

[1] Aditya Mahimkar, *IQ-OTHNCCD Lung Cancer Dataset*, 2021. Kaggle. Available at: https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset, Accessed: 2024-12-04.

[2] H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, ”Diagnosis of lung cancer based on CT scans using CNN,” in *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 2, pp. 022035, 2020.

[3] J. A. Bartholomai and H. B. Frieboes, ”Lung cancer survival prediction via machine learning regression, classification, and statistical techniques,” in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 632–637, 2018.

[4] S. Bharati, P. Podder, and P. K. Paul, ”Lung cancer recognition and prediction according to random forest ensemble and RUSBoost algorithm using LIDC data,” *International Journal of Hybrid Intelligent Systems*, vol. 15, no. 2, pp. 91–100, 2019.

[5] D. P. Kaucha, P. W. C. Prasad, A. Alsadoon, A. Elchouemi, and S. Sasikumaran, ”Early detection of lung cancer using SVM classifier in biomedical image processing,” in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pp. 3143–3148, 2017.