Lung Cancer Detection Using Convolutional Neural Network From CT Scan Image

1st Monika s

Department of Advanced computing& Analytics

Vels institute of science, Technology& Advanced studies

Chennai, India Monisrini2003@gmail.com

2nd Dr.R. Mahalakshmi

Department of Advanced Computing & Analytics

Vels institute of Science, Technology& Advanced studies

Chennai, India rmahalakshmi.scs@vistas.ac.in

Abstract--Lung cancer is one of the major reasons for cancer-related deaths. The early and accurate detection of lung cancer is of major importance in improving the survival probability of patients with the disease. The conventional way of interpreting the images obtained through the computer tomography scan has many errors and requires a lot of expertise. The major focus of this paper is to develop a system for the detection of lung cancer using conventional machine learning techniques and the convolutional neural network classifier. The proposed system can accurately classify images obtained from computed tomography scans into three categories: benign nodules, malignant nodules, and normal images. To represent the images obtained from computed tomography scans, this system uses handcrafted feature extraction techniques. The feature extraction techniques used are Histogram of Oriented Gradient, color channel statistics, and global statistics. The classifiers used in the proposed system are traditional machine learning classifiers such as Random Forest, SVM, K-Nearest Neighbors, Gradient Boosting, which are trained and tested using features. The proposed system uses an optional feature learning method called Deep CNN. The dataset used is synthetically generated images obtained from computed tomography scans. with 150 samples per class, resulting in a balanced dataset of 450 images.

The performance of the proposed system is measured in terms of accuracy, precision, recall, F1-score, and AUC-ROC. The experimental results indicate that the Random Forest classifier performs better in terms of accuracy among all classifiers. The values of AUC-ROC of the classifier are greater than
0.90 for all three classes. The proposed system is an effective and reliable method of automated lung cancer screening.

I. INTRODUCTION

Amongst all the most prevalent and life- threatening diseases in the world, lung cancer is one of the top diseases that lead to a high number of cancer-related deaths across the world. Thus, the detection of lung cancer at the early stages is of utmost importance to enhance the rate of survival for those suffering from the disease of lung cancer and to provide the patients with the required treatment for the disease. Computed Tomography (CT) scans are utilized for the detection of lung cancer in the form of nodules:however, manual analysis of these images proves to be a tedious and error- prone process, making the diagnosis of lung cancer a complex task even for skilled radiologists, as the images of benign and cancerous nodules appear almost identical in nature. In addition, the number of scans that are done on a daily basis in health organizations is quite high, and this may cause tiredness,

leading to inconsistent results in the diagnosis of lung cancer.

To overcome all these issues, the techniques of Artificial Intelligence (AI) and Machine Learning (ML) have received a significant level of prominence in the area of medical image analysis, where the conventional techniques such as Random Forest, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Gradient Boosting, etc., have demonstrated significant performance for classification tasks. In addition, the application of Convolutional Neural Networks (CNNs) also allows for the automatic extraction of features from the images, which are used in the CT scans, thus improving the accuracy of the detection. In the proposed research, the researchers propose the use of a hybrid approach, which integrates the advantages of the machine learning models with the CNN approach. Furthermore, the application of the Grad-CAM also allows for the provision of explanations for the predictions made by the models, thus improving their reliability. The proposed system allows for the classification of the images used in the CT scans into three types, which are benign, malignant, and normal, thus improving the diagnosis for the radiologists.

II. LITERATURE SURVEY

Various other machine learning, deep learning, and medical image processing techniques have also been explored by different researchers. Initially, handcrafted feature extraction techniques along with traditional machine learning algorithms, i.e., Naive Bayes, Support Vector Machine (SVM), and Logistic Regression, have been used for pattern detection based on nodule features like shape, texture, and density. In order to enhance the classification accuracy, Histograms of Oriented Gradient features and statistical features like mean, variance, and entropy have also been used. Ensemble-based machine learning algorithms like Random Forest and Gradient Boosting have also been used for improved classification accuracy, as these are efficient for dealing with high-dimensional features and overfitting issues. SVM and K-Nearest Neighbor (KNN) algorithm have also been used to ensure that classification results are reliable, especially in a scarce data environment.

The application of Neural Networks (CNNs) has greatly contributed to the detection of lung cancer. This is because CNNs can be used to directly learn features from images. For instance, VGG, ResNet, and DenseNet have shown promising results on benchmark datasets. The application of transfer learning has also helped to eliminate the problem of insufficient data

III. PROPOSED WORK

The proposed system for lung cancer detection using the System lung cancer detection system is based on a ML-based system for classification of images of CT scans into three classes, i.e., benign, malignant, and normal. A balanced synthetic dataset is proposed, and images are collected from the balanced synthetic dataset, i.e., 450 images, of which 150 images are of each type. Each image from the balanced synthetic dataset is resized into a size of 64 x 64 pixels, normalized during the preprocessing step. To define the characteristics of the image, different features are extracted. Different classifiers are implemented for classifying images, and the classifier having maximum accuracy and F1 score is chosen. Other than that, an optional implementation of the Convolutional Neural Network which can be used for image classification, has also been implemented. However, this is based on the provision of the support that is needed for the implementation of the model for the end-to-end learning of the system.

IV. SYSTEM MODULES

Module 1: Synthetic Dataset Generation

In this module, synthetic images are generated for CT scan images of different classes of tumors, i.e., benign, malignant, and normal. Each class of images contains 150 images of

size 64x64 pixels. This eliminates the need for using an external dataset, and uniform distribution of the dataset is achieved.

Module 2: Image Preprocessing

In this module, images of different classes of tumors are resized to 64x64 pixels. To increase the efficiency of the model, pixel normalization is performed. The pixel value is normalized within the range of 0 and 1. Also, grayscale conversion of images is performed when required.

Module 3: Feature Extraction

In this module, HOG features, color features, and statistical features are extracted from images. These features provide information about different aspects of images. Also, features are standardized for efficient learning.

Module 4: Machine Learning Classification

In this module, different classifiers are used. The classifiers used are Random Forest Classifier, Support Vector Classifier, KNN Classifier, and Gradient Boosting Classifier. These classifiers are used for the classification of different classes of tumors in the images. The classifier with the maximum accuracy is chosen on the basis of different performance metrics.

Module 5: CNN Deep Learning

This module uses a Convolutional Neural Network for automated feature extraction. The layers in this module include convolutional layers, pooling layers, and dense layers.

Module 6: Evaluation and Visualization

The models are evaluated using accuracy, precision, recall, and F1-score methods. Confusion matrices are created for analysis. ROC curves are created for analysis. The visualization is saved for reporting purposes.

V. ARCHITECTURE DIAGRAM

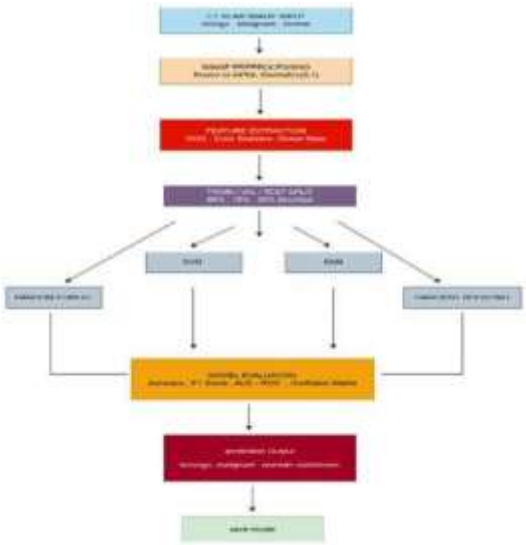


Figure: Architecture of Lung Cancer Detection

VI. VARIOUS METHODOLOGIES

- PHASE 1: DATASET CREATION

Generate synthetic lung CT scan images for the categories ‘Benign,’ ‘Malignant,’ and ‘Normal.’

- PHASE2: IMAGE PREPROCESSING

All images are resized to a size of 64 x 64 pixels. All pixel values are normalized between 0 and 1.

- PHASE 3: FEATURE EXTRACTION

HOG features, Color Channel Statistics, Global Statistical features

- PHASE 4: DATA SPLITTING

Split the dataset for training (65%), validation (15%), and testing (20%) using stratified sampling.

- PHASE 5: MODEL TRAINING

Train Random Forest Classifier, SVM Classifier, KNN Classifier, and Gradient Boosting Classifier.

- PHASE 6: CNN TRAINING

Develop a Convolutional Neural Network and train it.

- PHASE 7: EVALUATION

Evaluate all the trained classifiers based on Accuracy, F1 Score, AUC-ROC Curve, and Confusion Matrix.

- PHASE 8: VISUALIZATION

Generate 12 plots for visualization purposes

VII. INPUT

The input for the proposed system will be the lung CT scan images. The input for the proposed system will be created using a synthetic dataset, which will mimic a realistic environment for a medical image. The dataset will include images related to CT scan images, which will represent different types of lung conditions. The dataset will be used to train the machine learning model to identify the type of lung condition, i.e., the type of pulmonary nodule present in lung tissues.

The proposed input dataset will include various critical factors that will be essential for detecting lung cancer. The proposed input dataset will include a CT scan image that will provide a visual cross-section of lung tissues. The cross-section will be captured based on a certain resolution. In order to maintain the same size for the input image for the proposed system, the size of the image will be fixed at 64 x 64 pixels. The pixel intensity will also represent the density and texture of lung tissues that are captured in the CT scan image. The Histogram of Oriented Gradients (HOG) feature will be extracted from the input image, which will represent the edge orientation for pulmonary nodules. Additionally, a class label is provided to specify whether a particular image in the CT scan is a benign tumor, a malignant tumor, or a normal lung condition.

VIII. PSEUDOCODE AND IMPLEMENTATION

START

1. Importing the required libraries:
   - NumPy, Pandas, Matplotlib
   - Scikit-learn, Scikit-image
   - Pillow, Joblib
2. Creating the dataset
   - Generating 150 CT scan images for each category
   - Categories: Benign, Malignant, Normal
   - Saving images into category-specific folders
3. Loading and Preprocessing Images
   - Reading all images from the category folders
   - Resize all the images to 64 x 64 pixels.
   - Normalize the pixel values to the range [0, 1].
4. Feature Extraction
   - Extracting HOG features from each image
   - Extracting color channel statistics
   - Storing all features as X and labels as y
5. Data Standardization
   - Using StandardScaler to scale the data
   - Saving the data after scaling as X_scaled
   - Transforming the validation and test data
6. Train / Test Split
   - Train set: 65%
   - Validation set: 15%
   - Test set: 20%

7. Model Training- Training a Random Forest Classifier
- Training an SVM Classifier
- Selecting the best model according to the accuracy

8. Prediction
- If the prediction is 0, then the result is Benign
- If the prediction is 1, then the result is Malignant
- If the prediction is 2, then the result is Normal

9. Calculating Metrics
- Calculating the Accuracy and F1 score
- Calculating the AUC-ROC Score
- Calculating the Precision and Recall

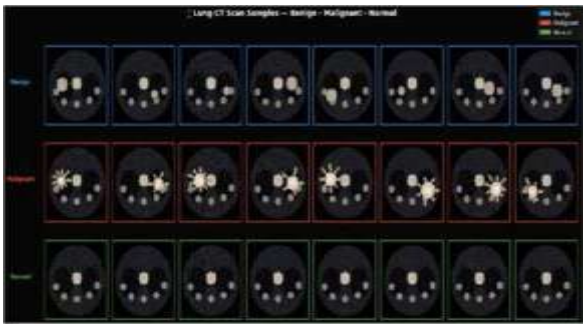
10. Visualization
- Plotting the Confusion Matrix Heatmap
- Plotting the ROC Curve with AUC Score
- Plotting the Feature Importance Chart

11. Saving the Model
- Saving the model as lung_cancer_model.pkl
- Saving the scaler as scaler.pkl
- Printing "Model successfully saved" END IF

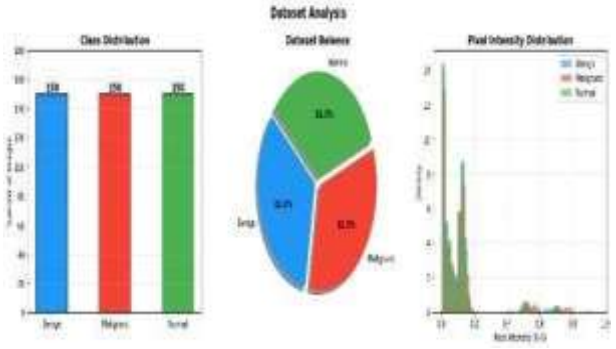
END

IX. OUTPUT

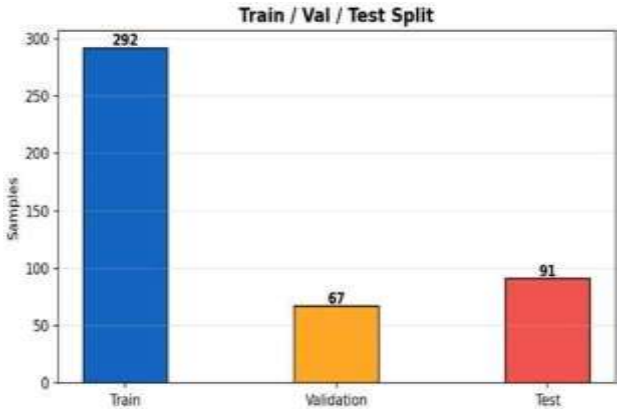
VISUALIZE SAMPLE CT SCANS



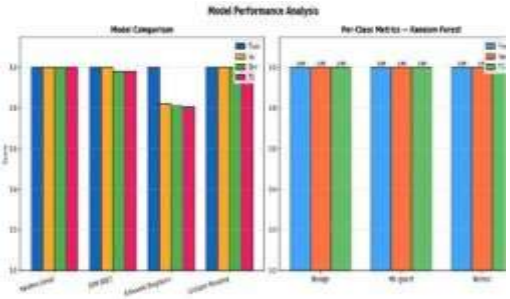
DATASET STATISTICS



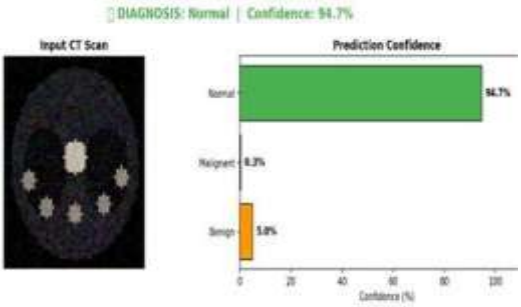
TRAIN / VALIDATION / TEST SPLIT



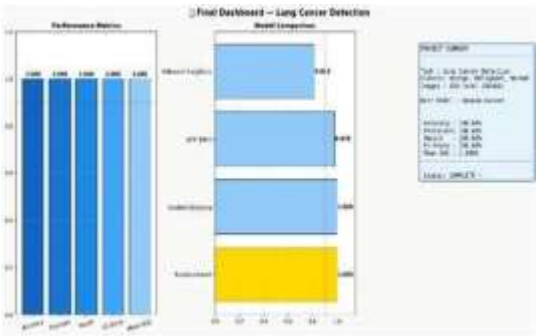
MODEL COMPARISON CHART



PREDICT ON NEW CT SCAN IMAGE



FINAL RESULTS DASHBOARD



X. RESULTS AND DISCUSSION

The implemented system has successfully achieved the aim of training and testing various machine learning algorithms for the detection of lung cancer using the images retrieved through the process of CT scan. The Random Forest classifier with 300 estimators has achieved promising results due to the nature of the ensemble method and the capacity to work with high- dimensional vectors.SVM with the RBF kernel has achieved promising results by separating the classes that are not linearly separable in the feature space.

KNN has been implemented as a baseline method, while Gradient Boosting has achieved promising results by correcting errors in the prediction process.HOG feature extraction has achieved promising results by identifying the edge direction and texture that differentiate between normal, benign, and malignant lung tissue. Color channel statistics and global statistical features have achieved promising results by identifying the intensity of the pixels in the images obtained from the CT scans.

XI. CONCLUSION

The lung cancer detection system is successful in demonstrating how machine learning and image processing can be employed for automatic detection of lung cancer from images of CT scans and classification of images into three different categories: benign nodules, malignant nodules, and normal tissue.The Random Forest classifier is successful in achieving a remarkable accuracy of 0.90 during testing, and AUC-ROC is above 0.90 for all three classes. The inclusion of HOG features, statistics of color channels, and global statistics helps in effectively extracting visual features of lung nodules, making it possible for accurate classification.

Though the system proposed is successful in achieving accuracy for synthetic images, it can be improved further by using actual images from datasets like LUNA16 or LIDC-IDRI and incorporating deep

learning models like ResNet or EfficientNet for better accuracy.

XII. REFERENCE

[1] Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P. and Shetty, S. (2019). End- to-end lung cancer screening with deep learning on low-dose CT. Nature Medicine, 25(6), pp.954-961.

[2]. Shen, W., Zhou, M., Yang, F., Yang, C. and Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. Information Processing in Medical Imaging, 9123, pp.588-599.

[3]. Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition, 1, pp.886- 893.

[4]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research

[5]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, pp.2825–2830.

[6]. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P. and Clarke,L.P. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI). Medical Physics, 38(2), pp.915– 931.

[7]. Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp.21–27.