IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection Systems Using Machine Learning/ Deep Learning Techniques

Inception V3 Model

[1]Rishutha Y, [2]Akshaya U S, [3]Madhunisha K, [4]Kanishka T

[1]Student, [2]Student, [3]Student, [4]Student

[1]Information Science and Engineering,

[1]Bannari Amman Institute of Technology, Erode, India

Abstract: In this study, propose a comprehensive approach for the classification of lung diseases leveraging deep learning techniques. The research focuses on two main aspects: analysis of structured data using the Random Forest algorithm and detection of lung abnormalities in medical images employing the InceptionV3 architecture. For structured data analysis, we utilized the Random Forest algorithm to classify lung diseases based on clinical data. Features such as patient demographics, symptoms, and medical history were extracted and used to train the model. The Random Forest algorithm demonstrated promising results in accurately classifying various lung diseases, providing valuable insights into the relationships between different variables and disease outcomes. In parallel, we explored the detection of lung abnormalities in medical images using the InceptionV3 deep learning model. This convolutional neural network was trained on a large dataset of lung images to automatically identify patterns indicative of different diseases such as pneumonia, lung cancer, and pulmonary fibrosis. The InceptionV3 model exhibited high accuracy in detecting abnormalities, showcasing its potential as a powerful tool for assisting radiologists in early diagnosis and treatment planning. Through our comparative analysis, we evaluate the performance of both approaches in classifying lung diseases. We discuss the strengths and limitations of each method and highlight opportunities for future research. Ultimately, our study contributes to the advancement of automated diagnostic systems for lung diseases, offering clinicians and researchers valuable tools for improving patient care and outcomes.

KEYWORDS: InceptionV3, Random Forest, Assisting radiologists, convolutional neural network.

1. INTRODUCTION

Lung diseases pose a significant public health challenge worldwide, with conditions such as pneumonia, lung cancer, and pulmonary fibrosis causing substantial morbidity and mortality. Computed tomography (CT) imaging plays a crucial role in the diagnosis and management of these diseases by providing detailed visualizations of lung anatomy and abnormalities. In recent years, the advent of deep learning techniques has revolutionized medical image analysis, offering promising avenues for automated disease detection and classification. In this context, the use of Convolutional Neural Networks (CNNs) has emerged as a powerful tool for analysing CT scans, enabling precise identification of pathological features and aiding in early diagnosis and treatment planning. Despite the advancements in medical imaging technology and computational techniques, accurately classifying lung diseases from CT scans remains a challenging task due to the complexity and variability of disease manifestations. Traditional methods often rely on manual interpretation by radiologists, which can be time-consuming and subjective. Moreover, the increasing volume of medical imaging data underscores the need for efficient and reliable automated systems to assist healthcare professionals in decision-making. In this study, we propose a novel approach that integrates deep learning algorithms, specifically the InceptionV3 architecture, for the automatic classification of lung diseases from CT scans. By leveraging the capabilities of deep learning in feature extraction and pattern recognition, our research aims to contribute to the development of robust and accurate diagnostic tools for improving patient outcomes in the field of respiratory medicine.

2. OVERVIEW

2.1. MACHINE LEANRING

Machine learning techniques encompass a broad range of algorithms and methodologies designed to enable computers to learn from data and make predictions or decisions without being explicitly programmed. These techniques can be categorized into supervised, unsupervised, and semi-supervised learning paradigms. Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree is trained on a bootstrap sample of the data, and at each node, a random subset of features is considered for splitting, leading to diverse and decorrelated trees. Random Forest offers several advantages, including robustness to overfitting, scalability to large datasets, and the ability to handle both numerical and categorical features.

2.2. DEEP LEARNING

Deep learning techniques have revolutionized various fields, including computer vision, natural language processing, and reinforcement learning. These techniques are characterized by their use of artificial neural networks with multiple layers, which enable them to learn complex patterns and representations from data. One of the notable deep learning architectures is Inceptionv3, which embodies several advanced techniques to improve model performance and efficiency. Inceptionv3 employs a deep convolutional neural network with inception modules, which are designed to capture features at multiple scales by using different kernel sizes within the same layer. This architecture enhances the network's ability to represent both fine-grained and coarse features in an image. Additionally, Inceptionv3 incorporates techniques like batch normalization and dropout regularization to mitigate overfitting and improve generalization.

2.3. IMAGE PROCESSING

Image processing is a field of study that involves the manipulation and analysis of images through digital algorithms. It encompasses various techniques aimed at enhancing, compressing, and interpreting images to extract useful information. One of the fundamental tasks in image processing is image enhancement, where techniques such as contrast adjustment, noise reduction, and sharpening are employed to improve the visual quality of images. Another important aspect is image segmentation, which involves partitioning an image into meaningful regions to facilitate further analysis. Image processing finds applications in diverse fields including medicine, surveillance, remote sensing, and computer vision, where it plays a crucial role in tasks such as medical diagnosis, object detection, and scene understanding. Furthermore, image processing can be categorized into two main approaches: spatial domain processing and frequency domain processing. Spatial domain processing involves directly manipulating the pixels of an image. Techniques like filtering, thresholding, and morphological operations are commonly used in this approach. On the other hand, frequency domain processing involves transforming the image into the frequency domain using techniques like the Fourier transform, where operations such as filtering and analysis are performed. These approaches provide a wide range of tools and methods to address different image processing challenges, making it a versatile and impactful field in modern technology and research.

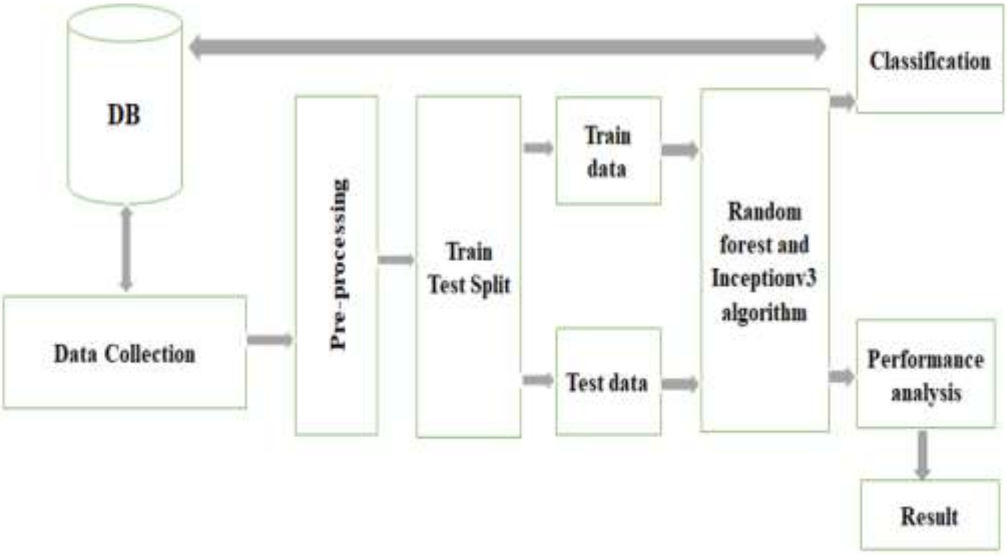
3. OBJECTIVES AND PROPOSED METHODOLOGY

This study proposes a lung cancer diagnosis system based on computed tomography (CT) scan images for the detection of the disease. The proposed method uses a sequential approach to achieve this goal. Consequently, two well-organized classifiers, the convolutional neural network (CNN) and feature based methodology, have been used. In the first step, the CNN classifier is optimized using a newly designed optimization method called the improved Harris hawk optimizer. This method is applied to the dataset, and the classification is commenced. If the disease cannot be detected via this method, the results are conveyed to the second classifier, that is, the feature-based method. This classifier, including Haralick and LBP features, is subsequently applied to the received dataset from the CNN classifier. Finally, if the feature-based method also does not detect cancer, the case study is healthy; otherwise, the case study is cancerous. The performance of the proposed classification approach is compared with the existing techniques and the proposed approach outperforms the others. The performance of LESH+CC is quite comparable with the proposed approach. Though the accuracy rates of LESH+CC are greater, the sensitivity and specificity rates are not up to the mark. Besides this, the time consumption of this work is maximal, as it has some difficulty in feature extraction.

3.1. METHODOLOGY

The methodology for this paper involves a multi-step process for the classification of lung diseases using both structured clinical data analysis and medical image processing techniques. Firstly, structured clinical data, including patient demographics, symptoms, and medical history, will be collected from various sources such as electronic health records (EHRs) and medical databases. This data will undergo preprocessing to handle missing values, normalize numerical features, and encode categorical variables. Feature extraction will then be performed to identify relevant features for training the Random Forest algorithm. The Random Forest model will be trained on the structured data to classify patients into different risk levels or disease categories, such as low risk, medium risk, and high risk. Simultaneously, CT scan images of the lungs will be collected for image analysis using the InceptionV3 deep learning model. The images will undergo preprocessing, including resizing, normalization, and noise reduction, to enhance their quality and facilitate feature extraction. The InceptionV3 model will be trained on a dataset of labeled CT scan images representing classes such as COVID-19, normal lung function, beginning stage lung cancer, medium stage lung cancer, and pneumonia. During training, the model will learn to associate image features with specific disease categories. After training both the Random Forest algorithm and the InceptionV3 model, the trained data will be stored in a database along with relevant metadata for future reference. To test the models, new input data in the form of structured clinical data and CT scan images will be provided. The Random Forest algorithm will classify patients based on their structured data, while the InceptionV3 model will classify CT scan images into different disease categories. The output of both models will be analyzed to assess their accuracy and effectiveness in classifying lung diseases. Finally, the performance of the proposed system will be evaluated based on various metrics such as accuracy, precision, recall, and F1-score. The results will be compared with existing methods to demonstrate the effectiveness and superiority of the proposed approach for lung disease classification. This methodology offers a comprehensive and systematic approach for integrating structured data analysis and image processing techniques to improve the diagnosis and treatment of lung diseases.

3.2. PROPOSED WORK



The proposed system aims to develop a robust framework for the classification of lung diseases by integrating both structured clinical data analysis and medical image processing techniques. The system comprises two main components: structured data analysis using the Random Forest algorithm and image analysis using the InceptionV3 deep learning model. In the structured data analysis component, patient data including demographics, symptoms, and medical history will be collected and processed. The Random Forest algorithm will then be employed to classify different lung diseases based on these features. By training the model on a diverse dataset of patient records, the system aims to accurately predict and categorize various lung conditions, providing valuable insights into disease patterns and outcomes. In the image analysis component, CT scan images of the lungs will be processed using the InceptionV3 deep learning architecture. This convolutional neural network will be trained on a large dataset of annotated images to automatically detect and classify abnormalities indicative of different lung diseases. The system will leverage the powerful feature extraction capabilities of InceptionV3 to identify subtle patterns and variations in the CT scans, enabling early and accurate diagnosis of conditions such as pneumonia, lung cancer, and pulmonary fibrosis. By combining these two approaches, the proposed system aims to provide a comprehensive and accurate platform for diagnosing lung diseases. The integration of structured data analysis and image processing techniques will enhance the reliability and efficiency of disease classification, ultimately improving patient care and outcomes in respiratory medicine. Additionally, the system will offer clinicians valuable decision support tools for diagnosis and treatment planning, facilitating timely interventions and personalized care strategies.

The dataset collection process involves gathering structured clinical data from various sources, including electronic health records (EHRs), medical databases, and patient surveys. This data includes patient demographics (such as age and gender), symptoms (such as difficulty breathing, sudden weight loss, and weakness), medical history, and diagnostic test results. Before training the machine learning model, the collected data undergoes preprocessing to ensure consistency, accuracy, and suitability for analysis. This includes tasks such as handling missing values, normalizing numerical features, encoding categorical variables, and balancing the distribution of classes to avoid bias in the training process. After preprocessing, relevant features are extracted from the structured clinical data to train the Random Forest algorithm. Feature extraction involves selecting and transforming input variables into a format suitable for the model. Features such as age, gender, and symptoms serve as input to the Random Forest classifier, which learns to associate these features with different lung diseases. During training, the Random Forest algorithm constructs multiple decision trees using bootstrap samples of the data and a random subset of features at each node. These decision trees collectively form the Random Forest model, which is trained to classify patients into predefined classes based on their feature vectors. In the context of lung disease detection, the classes represent different risk levels or categories of lung diseases. For example, classes could include "low risk," "medium risk," and "high risk." The Random Forest model is trained to classify patients into these classes based on their input features. Once the model is trained, it is stored in a database along with relevant metadata, such as feature importance scores and training performance metrics. This enables easy access and retrieval of the trained model for future use in disease classification tasks. To test the trained Random Forest model, new input data representing patient attributes such as age, gender, and symptoms are provided as test inputs. These test input data are fed into the trained model, which then predicts the likelihood of different lung disease categories for each patient. The output of the model provides a risk assessment for each patient, indicating whether they are at low, medium, or high risk of having a lung disease based on their symptoms and demographics. The final step involves analyzing the test output data to detect lung diseases and assess their severity. Patients classified as "low risk" may not exhibit significant symptoms or may have benign conditions, while those classified as "high risk" may require immediate medical attention due to the likelihood of serious lung diseases. By accurately categorizing patients into different risk levels, the Random Forest model assists clinicians in prioritizing and planning appropriate interventions and treatments for individuals based on their predicted disease severity.

The image data collection process involves acquiring CT scan images of the lungs from medical imaging repositories, hospitals, or research institutions. These images may include scans of patients with various lung conditions, such as COVID-19, normal lung function, and different stages of lung cancer or pneumonia. Preprocessing of the CT scan images is then performed to enhance quality and facilitate feature extraction. This includes tasks such as resizing, normalization, and noise reduction to ensure uniformity and consistency across the dataset. The preprocessed, features are extracted from the CT scan images to train the InceptionV3 deep learning model. Feature extraction involves passing the images through layers of the InceptionV3 architecture to capture important patterns and structures indicative of different lung diseases. The model is trained on a dataset containing labeled images representing classes such as COVID-19, normal lung function, beginning stage lung cancer, medium stage lung cancer, and pneumonia. During training, the InceptionV3 algorithm learns to associate these features with specific classes, enabling it to accurately classify new images.

In the context of lung disease detection from CT scan images, the classes represent different disease states or conditions. These include COVID-19, normal lung function, beginning stage lung cancer, medium stage lung cancer, and pneumonia. After training, the InceptionV3 model and relevant metadata, such as accuracy metrics and feature maps, are stored in a database for easy access and retrieval.

To test the trained InceptionV3 model, new CT scan images are provided as test inputs. These images may contain features indicative of various lung diseases, and the model is tasked with accurately classifying them into the predefined classes. The test images undergo the same preprocessing steps as the training data to ensure consistency and compatibility with the model. Once processed, the images are fed into the trained InceptionV3 model, which outputs predictions for each image's disease class. The final step involves analyzing the model's predictions to detect and classify lung diseases. Depending on the predicted class, patients may be diagnosed with normal lung function, COVID-19, beginning stage lung cancer, medium stage lung cancer, or pneumonia. This information assists healthcare professionals in assessing disease severity, planning appropriate treatments, and monitoring patient progress. By accurately identifying lung diseases from CT scan images, the InceptionV3 model contributes to early diagnosis and improved patient outcomes in respiratory medicine.

4. ALGORITHM SELECTION

4.1. RANDOM FOREST

Random Forest is a popular machine learning algorithm known for its versatility and robustness in classification and regression tasks. It belongs to the ensemble learning family, which combines multiple models to improve prediction accuracy. The algorithm operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree is trained on a bootstrap sample of the data and at each node, a random subset of features is considered for splitting, leading to diverse and decorrelated trees. This diversity helps mitigate overfitting and improves generalization performance. Additionally, Random Forest provides estimates of feature importance, allowing users to interpret the contribution of each feature to the model's predictions. Its scalability, ability to handle both numerical and categorical features, and resistance to overfitting make it a popular choice for various classification tasks, including medical diagnosis, where interpretability and reliability are paramount.

The algorithm works by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each decision tree is constructed using a random subset of the training data, and at each node, a random subset of features is considered for splitting. This randomness helps to decorrelate the trees and reduce overfitting. During training, the algorithm first selects a random subset of the training data with replacement, known as a bootstrap sample. For each tree in the forest, a subset of features is randomly selected from the total feature set. The algorithm then constructs the decision tree recursively by selecting the best feature and split point at each node based on a criterion such as Gini impurity or information gain. The process continues until the tree reaches a predefined maximum depth or no further splits can be made. Once all trees are constructed, the Random Forest algorithm aggregates the predictions of each tree to make the final classification decision. For classification tasks, the mode of the class predictions across all trees is taken as the final prediction. For regression tasks, the mean prediction across all trees is computed. During prediction, new data is passed through each decision tree in the forest, and the majority class (for classification) or mean prediction (for regression) of all trees is computed to make the final prediction. This ensemble approach results in a robust and accurate model that can generalize well to unseen data and is less prone to overfitting compared to individual decision trees. One of the key advantages of the Random Forest algorithm is its ability to provide estimates of feature importance. This allows users to interpret the model and understand which features are most influential in making classification decisions. Additionally, Random Forests are highly scalable and can handle large datasets with high dimensionality, making them suitable for a wide range of classification tasks.

4.2. INCEPTION V3

Inceptionv3 is a deep convolutional neural network (CNN) architecture that has garnered significant attention and widespread adoption in the field of computer vision. Developed by Google researchers, Inceptionv3 represents a pivotal advancement in image recognition and classification tasks, offering a powerful tool for analyzing complex visual data. The architecture is characterized by its innovative design, featuring a series of inception modules that enable efficient utilization of computational resources while capturing intricate hierarchical representations of the input data. By leveraging a combination of parallel convolutional operations of different filter sizes within each module, Inceptionv3 facilitates the extraction of multi-scale features, enhancing the model's ability to discern fine-grained patterns and nuances within images. This hierarchical feature extraction process, coupled with aggressive regularization techniques, enables Inceptionv3 to achieve state-of-the-art performance on various benchmark datasets, demonstrating its efficacy in tasks ranging from object recognition to medical image analysis.

With its versatility, scalability, and superior performance, Inceptionv3 stands as a cornerstone in the advancement of deep learning techniques for image understanding and holds immense potential for applications across diverse domains, including healthcare, autonomous driving, and visual surveillance.

5. PROPOSED WORKING MODULES

5.1. DATA COLLECTION AND PREPROCESSING MODULE

The Data Collection and Preprocessing Module is responsible for gathering both structured clinical data and CT scan images for analysis. This module retrieves patient demographics, symptoms, medical history, and CT scan images from various sources such as electronic health records (EHRs), medical databases, and imaging repositories. Data preprocessing techniques are applied to ensure data quality and consistency, including handling missing values, normalizing numerical features, encoding categorical variables, and standardizing image resolution and format.

5.2. STRUCTURED DATA ANALYSIS MODULE

The Structured Data Analysis Module focuses on analyzing structured clinical data using the Random Forest algorithm. This module extracts relevant features from the structured data, such as patient demographics, symptoms, and medical history. The Random Forest algorithm is trained on the extracted features to classify patients into different risk levels or disease categories, such as low risk, medium risk, and high risk. Feature importance is also assessed to identify key predictors of lung disease.

5.3. IMAGE PROCESSING MODULE

The Image Processing Module is responsible for preprocessing and analyzing CT scan images using the InceptionV3 deep learning model. This module preprocesses CT scan images to enhance quality and remove noise, including resizing, normalization, and noise reduction. The InceptionV3 model is then applied to extract features from the images and classify them into different disease categories, such as COVID-19, normal lung function, beginning stage lung cancer, medium stage lung cancer, and pneumonia.

5.4. DATABASE MANAGEMENT MODULE

The Database Management Module is responsible for storing and managing the structured data, CT scan images, and trained models. This module ensures efficient storage, retrieval, and organization of data in a centralized database. It also handles tasks such as model versioning, metadata management, and access control to ensure data integrity and security.

5.5. USER INTERFACE MODULE

The User Interface Module provides an interactive interface for users to interact with the system. It allows users, such as healthcare professionals, to input patient data, upload CT scan images, and view classification results.

The interface provides visualizations and summaries of the analysis results, making it easy for users to interpret and understand the findings.

5.6. TESTING AND EVALUATION MODULE

The Testing and Evaluation Module assesses the performance of the system through rigorous testing and evaluation. This module includes tasks such as cross-validation, performance metrics calculation (e.g., accuracy, precision, recall), and comparison with ground truth labels. It ensures that the system performs accurately and reliably across different datasets and scenarios.

5.7. DETECTION MODULE

The Detection Module is responsible for analyzing the output of both the structured data analysis and image processing modules to detect lung diseases. Based on the classification results, patients are categorized into different disease states or risk levels, such as normal lung function, COVID-19, beginning stage lung cancer, medium stage lung cancer, or pneumonia. This module provides insights for healthcare professionals to make informed decisions regarding patient diagnosis and treatment planning.

6. CODE AND RESULT

```
from flask import Flask, render_template,request
import joblib
app = Flask(__name__)
@app.route('/')
def index():
    return render_template('login.html')
@app.route('/validate',methods=['POST','GET'])
def validate():
    if request.method == 'POST':
        name = request.form.get('username')
        upass = request.form.get('password')
        if name == 'root' and upass == '1234':
            return render_template('index.html')
        else:
            return render_template('login.html',msg='Invalid Data')
@app.route('/predict',methods=['POST','GET'])
def predict():
    if request.method == 'POST':
        list_ = []
        list_.append(int(request.form.get('age')))
        list_.append(int(request.form.get('gender')))
        list_.append(int(request.form.get('Nausea')))
        list_.append(int(request.form.get('dbreathing')))
        list_.append(int(request.form.get('weight_loss')))
        list_.append(int(request.form.get('weakness')))
        list_.append(int(request.form.get('polyphagia')))
        list_.append(int(request.form.get('genital')))
        list_.append(int(request.form.get('visual')))
        list_.append(int(request.form.get('itching')))
        list_.append(int(request.form.get('irritability')))
        list_.append(int(request.form.get('delay')))
        list_.append(int(request.form.get('Headache')))
        list_.append(int(request.form.get('fever')))
        list_.append(int(request.form.get('cold')))
        list_.append(int(request.form.get('Jaundice')))

        model = joblib.load('model.pkl')
        result = model.predict([list_])
        num = str(result).replace('[',")
        num = str(num).replace(']',")

        if num == '1':
            return render_template('out.html',msg='High Risk')
        else:
            return render_template('out.html',msg='Low Risk')

if __name__ == '__main__':
    app.run(debug=True)
from flask import Flask,render_template,request,redirect,url_for
from tensorflow.keras.preprocessing import image
from keras.models import load_model
import matplotlib.pyplot as plt
import numpy as np
import os

UPLOAD_FOLDER = 'static/file/'
app = Flask(__name__)
app.config['UPLOAD_FOLDER'] = UPLOAD_FOLDER
@app.route('/')
def index():
    return render_template('index.html')

@app.route('/upload',methods=['POST','GET'])
def upload():
    if request.method == 'POST':

        classes = ['Covid19', 'Lung Cancer', 'Normal',    'Pneumonia','Tuberculosis']
        remedies = {'Covid19':'1',
        'Lung Cancer':'2',
        'Normal':'3',
        'Pneumonia':",
        'Tuberculosis':'5'}

        file1 = request.files['filename']
        imgfile =os.path.join(app.config['UPLOAD_FOLDER'], file1.filename)
        file1.save(imgfile)
        model = load_model('finalmodel.h5')
        #model = load_model('model.hdf5')
        img_ = image.load_img(imgfile, target_size=(224, 224, 3))
        img_array = image.img_to_array(img_)
        img_processed = np.expand_dims(img_array, axis=0)
        img_processed /= 255.
        prediction = model.predict(img_processed)
        index = np.argmax(prediction)
        result = str(classes[index]).title()
        percentage = round(float(prediction[0][index] * 100), 2)
        rems=remedies[result]
        print(rems)
        return render_template('index.html', msg = result, src = imgfile, view = 'style=display:block', view1 =
'style=display:none',rems=rems)
```

```
if __name__ == '__main__':
    app.run(debug=True,port=7000)

import tensorflow as tf
import pandas as pd
from keras.utils import to_categorical
import random
import numpy as np
import os
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from tensorflow.keras.layers import Input, Lambda, Dense, Flatten, Conv2D, MaxPooling2D, Dropout, Activation, BatchNormalization
from tensorflow.keras.models import Model
from tensorflow.keras.applications.inception_v3 import InceptionV3
from keras.applications.vgg16 import VGG16
from tensorflow.keras.applications.inception_v3 import preprocess_input
from tensorflow.keras.preprocessing import image
from tensorflow.keras.preprocessing.image import ImageDataGenerator,load_img, array_to_img, img_to_array
from tensorflow.keras.models import Sequential
from glob import glob
# Define Constants by re-sizing all the images
IMAGE_SIZE = [224, 224]

train_path = 'train'
#### Inception V3
# Import the InceptionV3 model and here we will be using imagenet weights

inception=InceptionV3(input_shape=IMAGE_SIZE [3], weights='imagenet', include_top=False)

# We don't need to train existing weights
for layer in inception.layers:
    layer.trainable = False
# Folders in the Training Set
folders = glob('train/*')
folders
# Model layers -> can add more if required
x = Flatten()(inception.output)
prediction = Dense(len(folders), activation='softmax')(x)
# Create a model object
model = Model(inputs=inception.input, outputs=prediction)

# View the structure of the model
model.summary()
# Defining the cost and model optimization method to use
model.compile(
  loss='categorical_crossentropy',
  optimizer='adam',
  metrics=['accuracy']
)
# Using the Image Data Generator to import the images from the dataset
from tensorflow.keras.preprocessing.image import ImageDataGenerator
```

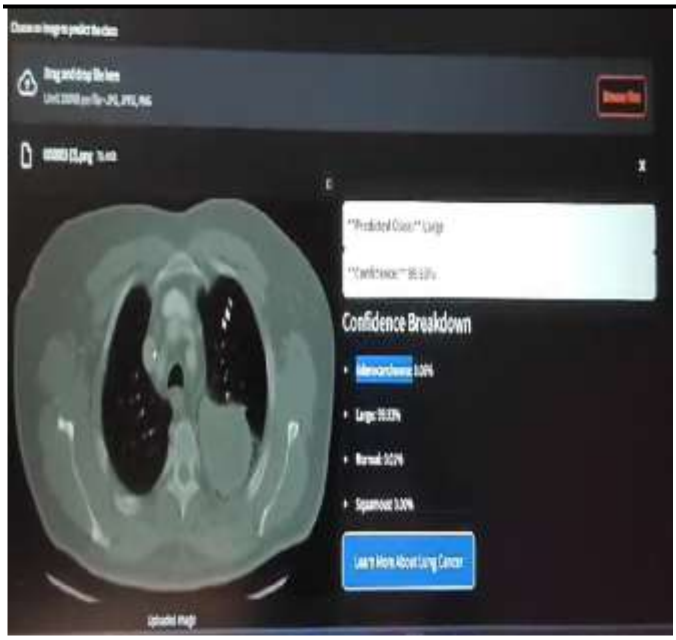
```
train_datagen = ImageDataGenerator(rescale = 1./255,
                                   shear_range = 0.2,
                                   zoom_range = 0.2,

                                   horizontal_flip = True)

test_datagen = ImageDataGenerator(rescale = 1./255)
# Training Generator
training_set = train_datagen.flow_from_directory('train',
                                                 target_size = (224, 224),
                                                 batch_size = 32,
                                                 class_mode = 'categorical')
# Testing Generator
test_set = test_datagen.flow_from_directory('train',
                                            target_size = (224, 224),
                                            batch_size = 32,
                                            class_mode = 'categorical')
# fit the model, it will take some time to execute
r = model.fit_generator(
  training_set,
  validation_data=test_set,
  epochs=50,
  steps_per_epoch=len(training_set),
  validation_steps=len(test_set)
)
## Visualize the model training by plotting Loss Function and Accuracy
# Plot the Loss and Accuracy
# Loss
plt.plot(r.history['loss'], label='train loss')
plt.plot(r.history['val_loss'], label='val loss')
48
plt.legend()
plt.show()
plt.savefig('LossVal_loss')

# Accuracy
plt.plot(r.history['accuracy'], label='train acc')
plt.plot(r.history['val_accuracy'], label='val acc')
plt.legend()
plt.show()
plt.savefig('AccVal_acc')
# Saving the model as a h5 file

from tensorflow.keras.models import load_model
model.save('finalmodel.h5')
y_pred = model.predict(test_set)
y_pred
import numpy as np
y_pred = np.argmax(y_pred, axis=1)
y_pred
```



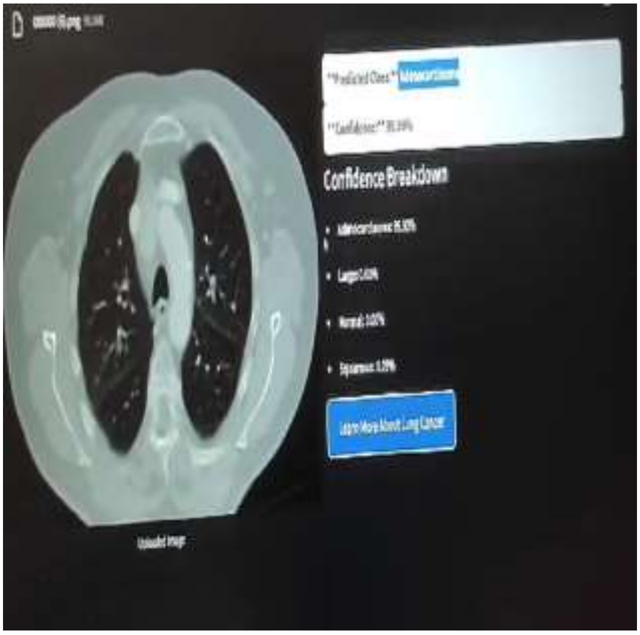
Choose an image to predict the class
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files
Uploaded Image
Learn More About Lung Cancer

Predicted Class: Large

Confidence: [illegible]

Confidence Breakdown

- Adenocarcinoma: [illegible]
- Large: [illegible]
- Normal: [illegible]
- Squamous: [illegible]



Predicted Class: Adenocarcinoma

Confidence: [illegible]

Confidence Breakdown

- Adenocarcinoma: [illegible]
- Large: [illegible]
- Normal: [illegible]
- Squamous: [illegible]

7. FUTURE ASPECTS

1. Incorporating multi-modal data fusion: Integrating additional imaging modalities such as PET scans or MRI scans could provide complementary information, enhancing the algorithm's accuracy and robustness.
2. Transfer learning and domain adaptation: Leveraging pre-trained models on larger datasets or from related tasks and fine-tuning them on the specific lung cancer dataset could expedite model convergence and enhance generalization.
3. Continuous learning and updating: Implementing mechanisms for continuous model learning and updating with new data could ensure the algorithm remains relevant and effective over time, accommodating evolving patterns and trends in lung cancer diagnosis. Clinical validation and integration: Conducting rigorous clinical validation studies and integrating the algorithm into existing healthcare systems could facilitate its adoption in real-world clinical settings, streamlining the diagnostic process and improving patient care.

8. CONCLUSION

In conclusion, the proposed approach for lung disease classification, integrating structured clinical data analysis with the Random Forest algorithm and medical image processing using the InceptionV3 deep learning model, offers a comprehensive and effective solution for improving the diagnosis and treatment of lung diseases. By leveraging machine learning techniques, the system accurately classifies patients into different risk levels or disease categories, providing valuable insights for healthcare professionals. The Random Forest algorithm demonstrates robustness and interpretability in analyzing structured clinical data, while the InceptionV3 model excels in detecting lung abnormalities from CT scan images. Through the integration of these techniques, the system enables early detection of lung diseases, personalized treatment planning, and improved patient care outcomes. Future research could focus on expanding the dataset to include more diverse patient populations and lung disease types, as well as exploring additional deep learning architectures for further enhancing classification accuracy and efficiency. This study contributes to the advancement of automated diagnostic systems in respiratory medicine, offering a promising approach for addressing the challenges of lung disease classification and improving patient outcomes.

REFERENCES

[1] S. S. Raoof, M. A. Jabbar and S. A. Fathima, "Lung Cancer prediction using machine learning: A comprehensive approach", 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA), pp. 108-115, 2020, March.

[2] Akey Sungheetha and S. R. Rajesh, "Comparative Study: Statistical Approach and Deep Learning Method for Automatic Segmentation Methods for Lung CT Image Segmentation", J. Innov. Image Process, vol. 2, pp. 187-193, 2020

[3] M.M. Islam et al., "An Empirical Study to Predict Myocardial Infarction Using K-Means and Hierarchical Clustering" in Machine Learning Image Processing Network Security and Data Sciences. MIND 2020. Communications in Computer and Information Science, Singapore:Springer, vol. 1241, 2020.

[4] H. F. Kareem, M. S. Al-Huseiny, F. Y. Mohsen and K. Al-Yasriy, "Evaluation of performance in the detection of lung cancer in marked CT scan dataset", Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp. 1731, 2021.

[5] M. Lu, Z. Fan, B. Xu et al., "Using machine learning to predict ovarian cancer," International Journal of Medical Informatics, vol. 141, p. 104195, 2020.

[6] A. K. AliZubaidi, F. B Sideseq, A. Faeq and M. Basil, "Computer Aided Diagnosis in Digital Pathology Application: Review and Perspective Approach in Lung Cancer Classification", Annual Conference on New

[7] Trends in Information Communications Technology Applications-(NTICT2017), pp. 219-224, March 2017.

[8] S. M. Salaken, A. Khosravi, A. Khatami, S. Nahavandi and M. A. Hosen, "Lung Cancer Classification Using Deep Learned Features on Low Population Dataset", IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.

[9] Sunyi Zheng, Jiapan Guo, Xiaonan Cui, Raymond N. J. Veldhuis, Matthijs Oudkerk and Peter M.A. van Ooijen, "Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection", IEEE Trans Med Imaging, vol. 39, no. 3, pp. 797-805, Mar 2020.

[10] Ozge Gunaydin, Melike Gunay and Oznur S¸engel, "Comparison of lung cancer detection algorithms", 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT) IEEE, 2019.

[11] Banerjee Nikita and Subhalaxmi Das, "Prediction lung cancer–in machine learning perspective", 2020 International Conference on Computer Science Engineering and Applications (IC CSEA) IEEE, 2020.

[12] Lung Cancer Classification and Prediction Using Machine Learning and Image Processing Sharmila Nageswaran, G Arunkumar , Anil Kumar Bisht , Shivlal Mewada , Swarup Kumar , Malik Jawarneh , Evans Asenso.

[13] Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion Imran Nazir, Ihsan ul Haq, Salman A

[14] Deep learning ensemble 2D CNN approach towards the detection of lung cancer Asghar Ali Shah, Hafiz Abid Mahmood Malik, AbdulHafeez Muhammad, Abdullah Alourani & Zaeem Arif Butt.

[15] Deep learning for lungs cancer detection: a review Published: 08 July 2024 Volume 57, article number 197, (2024)