המכללה
האקדמית להנדסה
בראודה
בכרמיאל

Capstone Project Phase A

24-1-R-1

Estimation Of Smoking Associated Damage Based On Nuclear Lung Images

Supervisor: Zeev Frenkel

Tareq Sleiman

Hani Zahran

2024

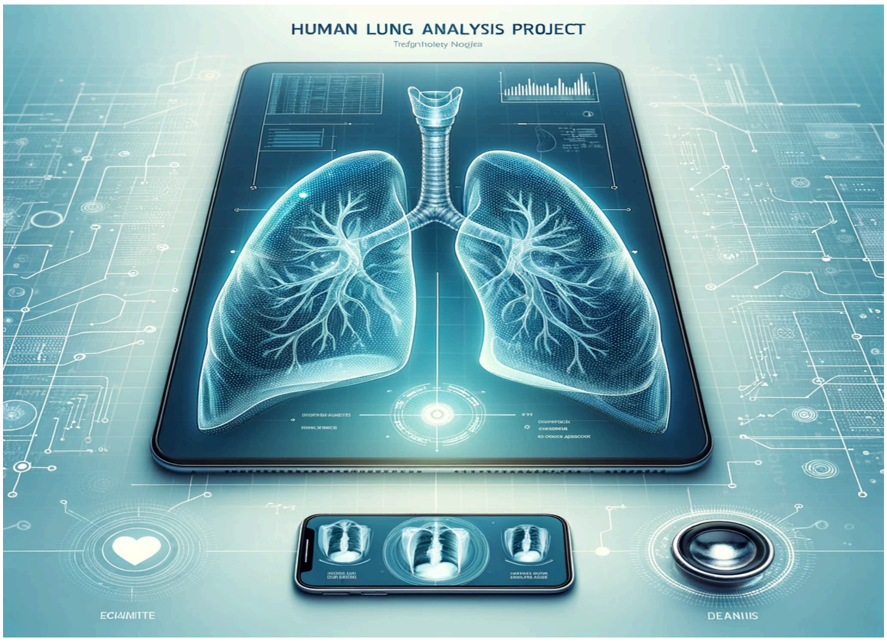


Table of contents

Abstract. In the dynamic intersection of healthcare technology and artificial intelligence, our project introduces an advanced software application designed to enhance the diagnostic accuracy and efficiency in detecting lung diseases attributable to smoking. Utilizing cutting-edge Convolutional Neural Networks (CNNs)[1], our tool capitalizes on deep learning algorithms to analyze lung imagery systematically. The application aims not only to differentiate between the lungs of smokers and non-smokers but also to identify specific conditions such as Chronic Obstructive Pulmonary Disease (COPD)[20], Emphysema, Lung Cancer[19], Bronchitis, and Pulmonary Fibrosis. This project leverages a rich dataset of annotated lung images from public medical repositories, ensuring a robust model training that captures subtle pathological features. Our CNN[1] model, developed with frameworks like TensorFlow and PyTorch and enhanced by GPU-accelerated computing, promises real-time, precise medical insights, thereby reducing the reliance on invasive procedures and subjective interpretations. By integrating seamlessly into existing medical infrastructures and adhering to stringent data protection standards, the software stands to significantly impact clinical workflows, offering a proactive tool in the fight against smoking-related illnesses. This initiative not only addresses a critical gap in current diagnostic processes but also sets a new standard in the application of AI in medical diagnostics, emphasizing user-friendliness and scalability.

1. Introduction

In the ever-evolving landscape of healthcare technology, our project sets out to marry the realms of artificial intelligence and medical diagnostics. Focusing on diseases caused by smoking, we aim to develop a groundbreaking software application capable of discerning crucial information from medical imagery that is meant to detect lung illness.

Project Overview

Firstly, we seek to create a tool that can distinguish between the lungs of smokers and non-smokers. Secondly, we endeavor to equip this tool with the ability to identify potential illnesses like Chronic obstructive pulmonary disease (COPD), Emphysema, Lung cancer, Bronchitis, And Pulmonary fibrosis. This dual-purpose application holds immense promise in streamlining diagnostics and fostering proactive healthcare.

Working principle

At its core, our software will operate on the principle of image processing by deep learning, a subset of machine learning that mimics the human brain's neural networks. By exposing the system to a diverse dataset of lung images, we enable it to autonomously learn and recognize patterns, empowering it to make informed decisions about the health and smoking history of an individual based on their lung scans. We plan to use a CNN model trained on a big dataset of real X-ray lung images with supplied information including marking of manually detected damage. This method was selected based on our education in college and consulting with several experts in the field.

Real-world Application

Once trained, our CNN model becomes a powerful tool for analyzing lung images. By inputting a scan, the software can swiftly provide insights into the individual's smoking habits and raise flags for potential health issues. This real-time analysis holds the potential to expedite diagnostics and inform timely medical interventions.

Implementation Possibilities

Alignment with Business Needs: Our software solution is designed to meet the critical needs of the medical and diagnostic industry by providing a non-invasive, accurate, and efficient method for detecting lung-related diseases. To ensure alignment with business needs, our application will be adaptable to different healthcare settings, such as hospitals, clinics, and diagnostic centers. Key requirements include compatibility with existing medical imaging hardware, scalability to handle varying patient loads, and customization options that allow for integration with other healthcare management systems. This alignment will enable healthcare providers to enhance diagnostic accuracy, reduce costs associated with late-stage treatments, and improve patient outcomes.

Technology and Engineering Approaches: Our project leverages state-of-the-art technologies in image processing and machine learning. We are utilizing advanced deep learning frameworks such as TensorFlow and PyTorch to develop and train our CNN models. The use of GPU-accelerated computing will significantly enhance the training speed and performance of our models. Additionally, we are exploring cloud-based solutions to ensure the scalability and accessibility of our application. These technologies are crucial for handling large datasets of lung images and for providing real-time analysis capabilities.

Algorithmic Approaches: We plan to implement several sophisticated algorithmic approaches, primarily focusing on Convolutional Neural Networks (CNNs)[1] for their proven effectiveness in image recognition tasks. Complementary algorithms such as Support Vector Machines (SVMs)[3] and Random Forests may also be integrated to enhance classification[4] accuracy. These algorithms are selected for their ability to learn complex patterns in imaging data, which is critical for identifying subtle indicators of lung disease. We will assess the advantages of each algorithm in terms of accuracy, speed, and computational efficiency, and address potential challenges such as overfitting and data bias.

Engineering Requirements: The document should state that the software needs to be highly functional and user-friendly. It should be designed to perform effectively and efficiently, meeting the needs of its users with a straightforward and intuitive interface. This ensures that users can operate the software easily without extensive training, enhancing their overall experience and satisfaction.

Proposed Solution and Its Benefits

Our project aims to develop artificial intelligence-based software that leverages deep learning to analyze medical images, specifically lung scans. By using Convolutional Neural Networks (CNNs)[1], our application will be able to differentiate between the lungs of smokers and non-smokers and identify early signs of lung diseases. This technology will:

- Enhance the accuracy and speed of lung disease diagnosis.

- Reduce the need for invasive diagnostic procedures.
- Provide a tool that is less dependent on individual radiologist expertise, thereby standardizing diagnosis processes.

Stakeholders and How They Benefit

The primary beneficiaries of this solution will be healthcare providers and patients. Healthcare providers will gain a powerful tool that aids in the quick and accurate diagnosis of lung conditions, which can be integrated into existing medical imaging systems. Patients will benefit from earlier and more accurate diagnoses, potentially leading to improved treatment outcomes. Additionally, healthcare policymakers can utilize such technologies to reduce healthcare costs associated with the treatment of advanced lung diseases.

Document Structure Overview

The remainder of this document will cover several key areas:

1. Background and Related Work: This section will delve into prior research and existing technologies in the realm of medical imaging and AI, highlighting the advancements and limitations of current methodologies.
2. Deep Learning in Action: We will explain the technical foundations of deep learning, particularly CNNs[1], and discuss our methodology for dataset augmentation and model training.
3. Algorithms: Description of the algorithmic approaches we plan to use, including a blend of CNNs[1], SVMs[3], and possibly other machine learning models.
4. Expected Achievements: We will outline the specific metrics and targets for our application, from accuracy in lung differentiation to user-friendly interface considerations.
5. Unique Features and Research/Engineering Challenges: Here we present innovative aspects of our project and the challenges we anticipate, including handling diverse data and ensuring real-time processing capabilities.
6. Preliminary Software Engineering Documentation: Overview of our development process, from initial model preparation to future plans for code development and system testing.

2. Background and Related Work

Why this project ?: We chose this project to address the significant health challenges posed by smoking-related lung diseases, which are among the leading causes of death globally. This project offers an opportunity to innovate in medical diagnostics by integrating advanced AI and deep learning technologies to improve early detection and treatment outcomes. It also fulfills a growing market need for non-invasive, accurate diagnostic tools in healthcare, providing us with a platform for significant professional and academic growth. This intersection of technology and health care, coupled with the potential for substantial societal impact, makes this project especially compelling.

2.1 Existing Solutions

Currently, the diagnosis of smoking-related lung diseases primarily relies on physical examinations, patient history, and conventional imaging techniques like X-rays and CT scans. These methods often require further validation through invasive procedures such as biopsies. Moreover, the interpretation of these images heavily depends on the expertise and subjective judgment of radiologists, which can lead to variability in diagnosis accuracy. Now we shortly review some of related works:

1. Google DeepMind – AI for Breast Cancer Analysis:

Google DeepMind developed an AI system that surpassed human radiologists in detecting breast cancer from X-rays. Although focused on breast rather than lung cancer, the project similarly utilizes deep learning techniques to analyze medical images and identify disease markers more accurately than traditional methods.

2. Lunit INSIGHT for Chest Radiography:

This is a suite of AI-powered diagnostic tools developed by Lunit, a South Korean medical AI company. Lunit INSIGHT specifically focuses on chest radiographs and is capable of detecting major thoracic abnormalities, including lung cancer, tuberculosis, and pneumonia with high accuracy. It's an exemplary case of applying CNNs in a real clinical setting for lung disease diagnostics.

3. IBM Watson Health Imaging:

IBM Watson Health uses AI to enhance imaging diagnostics across various conditions, including lung diseases. Their AI tools help radiologists interpret medical images more efficiently and with greater precision, showcasing how AI integration can improve diagnostic workflows and patient outcomes in lung health.

4. PathAI:

PathAI is using machine learning to assist pathologists in diagnosing and researching diseases from histopathology images. While their initial focus was not exclusively on lung diseases, they've ventured into areas that include lung pathology to improve diagnostic accuracy through automated, intelligent analysis.

5. RADLogics' AI-Powered Solutions for COVID-19:

During the COVID-19 pandemic, RADLogics provided AI-driven solutions to detect and quantify COVID-19 indications in chest CT scans. Their technology helped in assessing lung involvement and sped up the radiological workflow, which is critical in managing respiratory illnesses including those exacerbated by smoking.

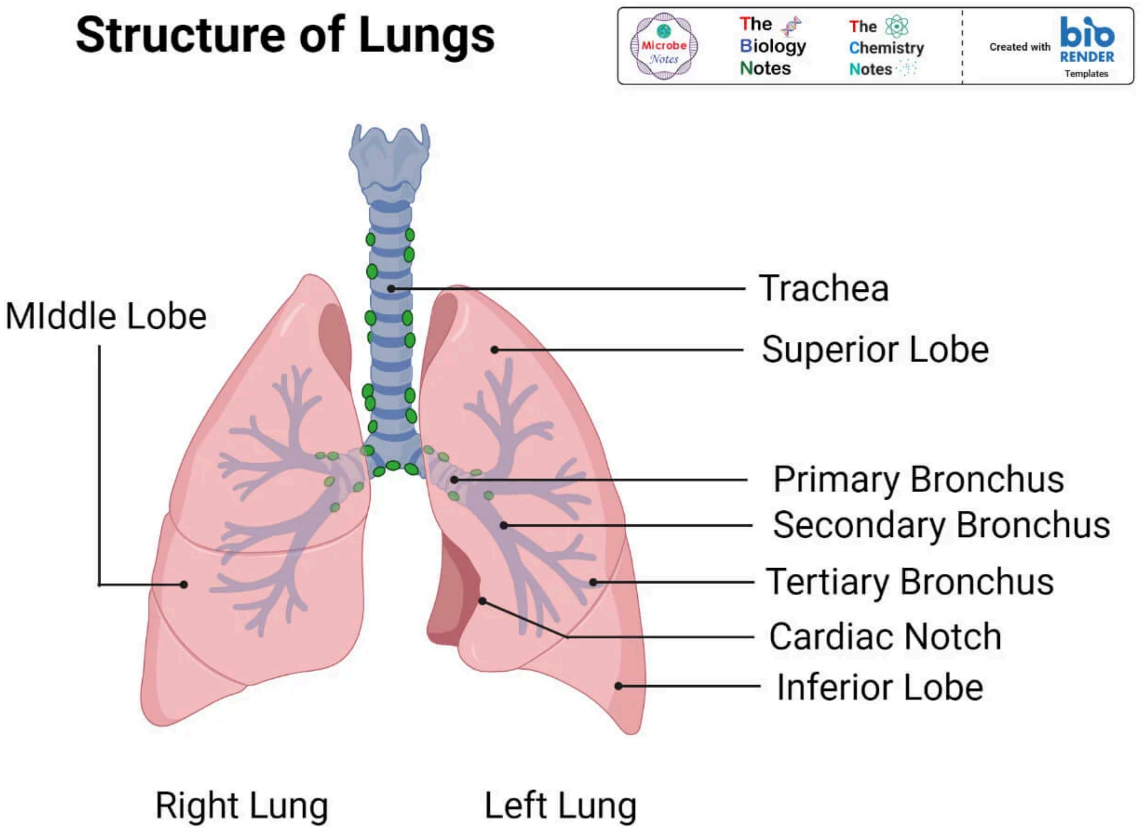
While we recognize that there are numerous projects addressing similar themes in the realm of medical imaging and AI diagnostics, our initiative is distinct in its targeted focus and execution. Our project aims to develop a specialized software that detects more lung illnesses related to smoking which some previous projects didn’t approach but also incorporates specific enhancements to distinguish it from existing solutions.

We are committed to improving the accuracy of detection for subtle lung damage caused by smoking, which may have been overlooked by earlier models. Additionally, a key aspect of our project is to enhance user-friendliness; we are designing an interface that is intuitive and straightforward, enabling healthcare professionals to use our software seamlessly within their existing workflows. This approach not only makes our tool more accessible but also increases its practical utility in clinical settings, ensuring that our software delivers real value by improving diagnostic processes and patient care outcomes.

2.2 Lungs of human

Structure of Lungs

The respiratory system is categorized into two main sections: the upper and lower respiratory systems. The upper respiratory system starts at the nose and nostrils, extends to the larynx, and concludes at the trachea. On the other hand, the lower respiratory system consists entirely of the lungs. Within the lungs, key components such as the trachea, alveoli, bronchi, and pleura each perform distinct functions.[16]



Figure[1]: Structure of Lungs.[16]

Location of Lungs [16]

- The lungs are located on either side of the chest, positioned within the thoracic cavity above the diaphragm—one on each side of the mediastinum. Encased in a pleural sac, they are shielded by the rib cage.
- The lungs are designated as the left and right lungs, corresponding to the sides of the body on which they are situated.

Anatomically lung presents with four features [16]

1. Apex: This is the upward-directed conical end of the lung.(Figure 1)
2. Base: This broader end is directed downwards. (Figure 1)
3. Borders:
    - The anterior margin is a thin edge that faces forward.
    - The posterior margin is a rounded border facing inward.
    - The inferior margin is a semilunar-shaped edge that divides the coastal surface from the medial surface.
4. Surfaces:
    - The coastal surface is convex and directed outward.
    - The medial surface is flat and also directed outward.

Chest X-ray images of lungs

Chest X-rays generate images that encompass the heart, lungs, airways, blood vessels, and the bones of the chest and spine. They can also detect fluid in or around the lungs or air surrounding a lung. When you visit your doctor or an emergency room due to chest pain, a chest injury, or difficulty breathing, a chest X-ray is often performed. This imaging helps your doctor assess whether you are experiencing issues such as heart problems, a collapsed lung, pneumonia, broken ribs, emphysema, cancer, or other potential conditions. For ongoing monitoring of a medical condition, some individuals undergo a series of chest X-rays over time to observe if their health issue is improving or worsening.

A chest X-ray can reveal many things inside your body, including:

1. **The condition of your lungs:** Chest X-rays can detect cancer, infection or air collecting in the space around a lung, which can cause the lung to collapse. They can also show chronic lung conditions, such as emphysema or cystic fibrosis, as well as complications related to these conditions.[17]
2. **Heart-related lung problems:** Chest X-rays can show changes or problems in your lungs that stem from heart problems. For instance, fluid in your lungs can be a result of congestive heart failure.[17]
3. **Postoperative changes:** Chest X-rays are useful for monitoring your recovery after you've had surgery in your chest, such as on your heart, lungs or esophagus. Your doctor can look at any lines or tubes that were placed during surgery to check for air leaks and areas of fluid or air buildup.[17]

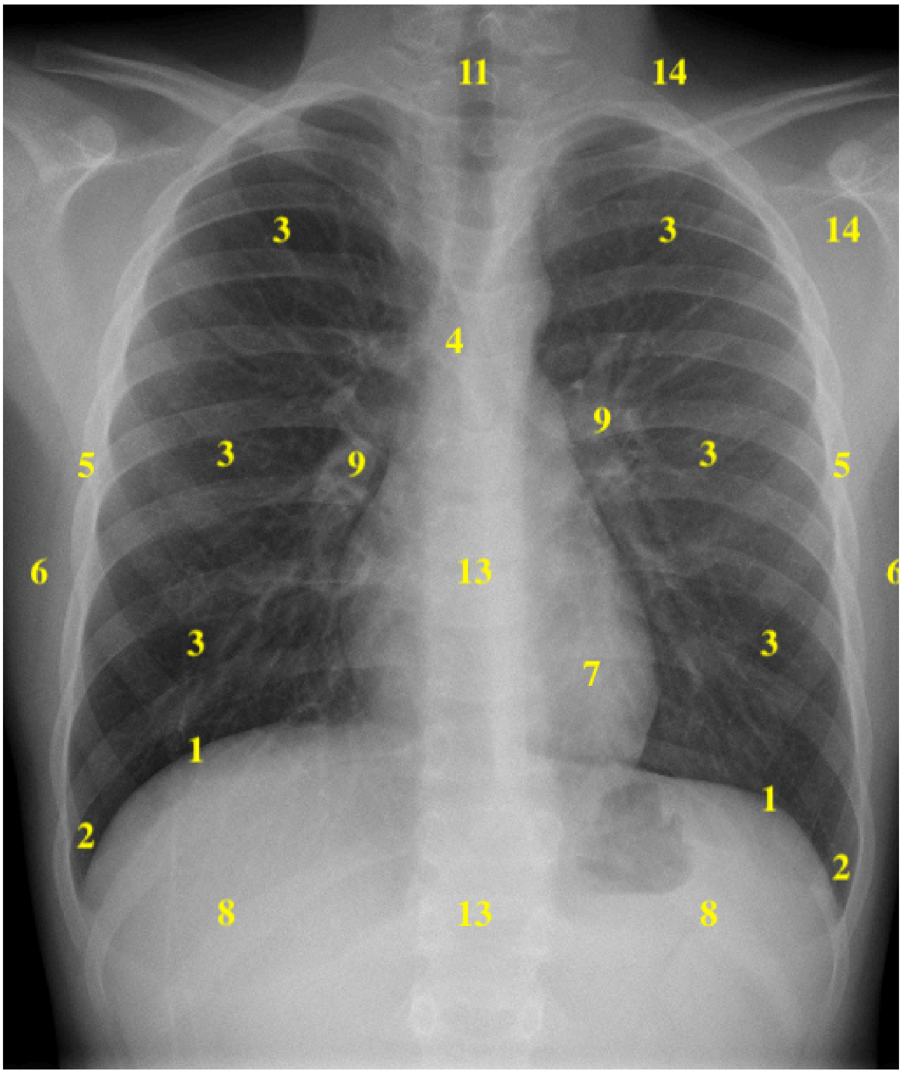


Figure [2]: Chest x-ray[17]

1	Hemidiaphragms	8	Below Hemidiaphragms
2	Costophrenic Angles	9	Hila
3	Zones of the Lungs	10	Anterior Mediastinum and Sternum
4	Carina and Bronchi	11	Trachea
5	Pleura and Ribs	12	Posterior Mediastinum
6	Chest Wall	13	Spine
7	Cardiac Silhouette	14	Soft Tissue of the Axilla and Lower Neck

2.3 The main observed features on X-ray lung images what we try to search

1. **Differentiate between the lungs of smokers and non-smokers**: In smokers, chest X-rays can show more pronounced signs of damage due to smoking-related lung diseases such as emphysema, chronic obstructive pulmonary disease (COPD), and increased likelihood of lung cancer. Non-smokers, particularly those affected by air pollution or occupational hazards, may show similar patterns in COPD but generally have less severe radiographic appearances. [18][19] (Figure 3)

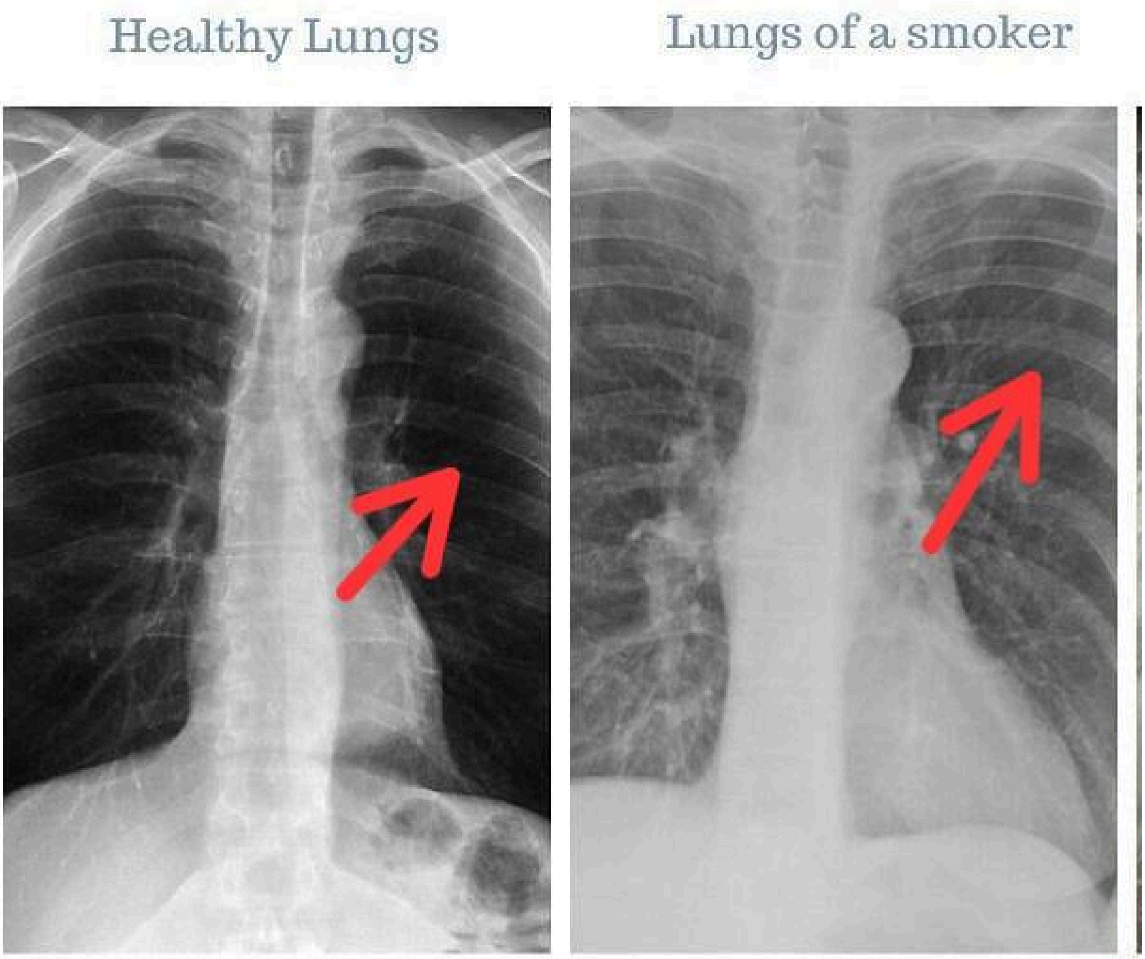
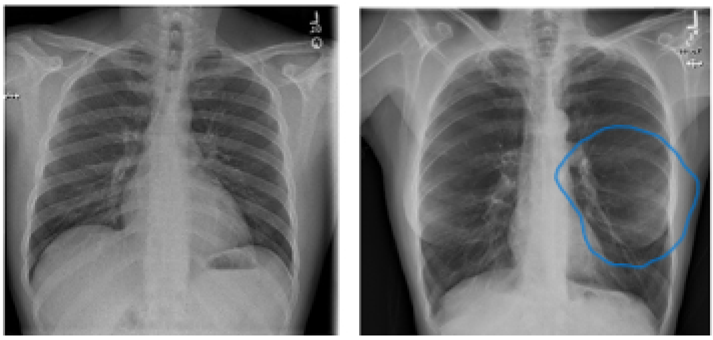


Figure [3]: Difference between smokers and non-smokers lungs

2. **Chronic Obstructive Pulmonary Disease (COPD)**: In X-rays, COPD can be indicated by signs of lung hyperinflation, a flattened diaphragm, and possibly an increase in the size of the pulmonary arteries, suggesting pulmonary hypertension. Bullae might also be visible, presenting as large air-filled spaces. [20](Figure 4)



Figure[4]: Difference between normal lungs and lungs with COPD illness

3. **Emphysema**: This appears as areas of decreased lung density, over-distended air sacs, and sometimes visible bullae. Emphysema caused by alpha-1 antitrypsin deficiency typically affects the lower parts of the lungs.[20][21](Figure 5)

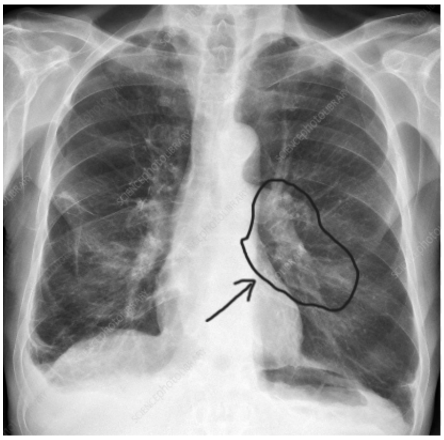
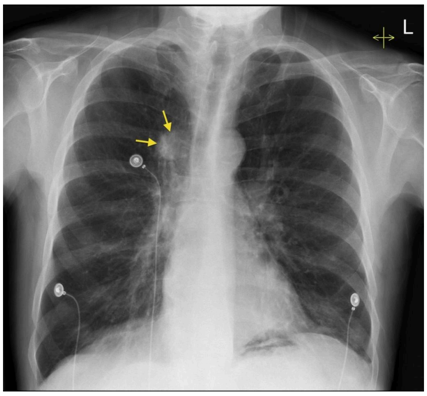


Figure [5]: Example of x-ray image of **Emphysema** illness

4. **Lung Cancer**: X-rays might reveal masses or nodules, and there can be signs of lung collapse or air surrounding the lungs if a tumor blocks an airway. Advanced cases might show signs of metastases to other areas like bones or the liver.[19](Figure 6)



Figure[6]: Example of x-ray image of **Lung Cancer** illness

2.4 Dataset Augmentation

Our model's effectiveness heavily depends on a comprehensive and diverse dataset. We are compiling a dataset that includes a wide range of lung X-ray and CT images from both smokers and non-smokers. This collection is sourced from public medical image repositories like the National Institutes of Health (NIH) Clinical Center Chest X-Ray dataset and the Lung Image Database Consortium (LIDC). By including varied stages of lung pathologies, we ensure our model is robust and reliable across different clinical scenarios.

Dataset:

- **NIH Chest X-Ray Dataset**

Our project is focused on analyzing nuclear lung images using deep learning, the following dataset could be highly relevant to our project:

NIH Chest X-ray Dataset (ChestX-ray14): This dataset comprises 112,120 frontal-view X-ray images from 30,805 patients. It includes labels for fourteen different thorax disease categories such as Atelectasis, Consolidation, and Emphysema, among others. These images and their diverse pathological labels could be invaluable for training models to recognize various lung conditions, including those related to smoking.You can access the NIH Chest X-ray Dataset directly through the NIH repository.[22]

2.5 Deep Learning in Action

Understanding Deep Learning: In our project, deep learning represents a transformative approach to medical image analysis, moving from manual feature extraction to automated, sophisticated pattern recognition. This shift enables our model to detect subtle and complex features in lung images that are often imperceptible to human observers, a critical capability for identifying early signs of diseases associated with smoking.

Why Choose CNN for This Project?

1. **Feature Detection:** CNNs are exceptionally good at detecting features in images due to their hierarchical architecture. They can learn to identify important features at various levels, from edges and textures at lower levels to more complex patterns at higher levels.[1]
2. **Automatic Feature Learning:** Unlike traditional methods that require manual feature extraction, CNNs automatically learn the best features for the task during the training process. This capability is crucial for medical images where subtle features can be critical indicators of disease.[5]
3. **Performance:** CNNs have demonstrated superior performance in image classification tasks, including medical imaging. Their ability to generalize from training data to new, unseen images makes them ideal for diagnosing lung diseases from X-rays or CT scans.[5]

4. **Adaptability:** The CNN architecture can be easily adapted and fine-tuned for specific tasks, such as distinguishing between different types of lung diseases or assessing the severity of disease from imaging data.[5]

Convolutional Neural Networks (CNNs)

We chose Convolutional Neural Networks (CNNs) for their proven effectiveness in image recognition tasks, particularly for processing and learning from image data. CNNs employ a hierarchical, layered structure that enables them to detect and interpret complex features at multiple scales within the images.

Model Structure

Our CNN architecture is designed to efficiently process medical imaging data:

1. **Input Layer:** Accepts raw input of lung images.
2. **Convolutional Layers:** These layers use various filters to detect different features in the images, such as edges and textural patterns.
3. **Activation Layers:** Introduce non-linearity, allowing the model to learn complex patterns.
4. **Pooling Layers:** Reduce the spatial dimensions of the image data, helping to make the model more computationally efficient.
5. **Fully Connected Layers:** Deep layers that integrate the learned features from previous layers to make final predictions.
6. **Output Layer:** Outputs the classification results, such as distinguishing between healthy, smoker, and diseased lungs.

Model Training

The training process involves several key phases:

1. **Feature Learning:** The model learns to identify essential features in the lung images that correlate with various smoking-related diseases.
2. **Optimization:** We employ advanced optimization algorithms like Adam or SGD (Stochastic Gradient Descent) to effectively minimize the loss function, which quantifies the discrepancies between the model’s predictions and the actual data.
3. **Validation:** Using a separate set of images, we continuously validate the model's accuracy and adjust parameters to avoid overfitting and ensure generalizability to new, unseen images.

3. Algorithm

For our software project focused on analyzing lung images to detect smoking-related diseases, utilizing a Convolutional Neural Network (CNN) would be highly suitable due to its proven effectiveness in image recognition tasks, particularly in medical imaging. Below is a basic outline of the algorithm and pseudocode for our software, along with an explanation of why CNNs are chosen for this application:

Algorithm Overview

1. **Data Preprocessing:**
    - Load the lung images.
    - Normalize the images to a standard scale.
    - Augment the dataset to improve model robustness (e.g., rotations, translations).
2. **Model Construction:**
    - Define a CNN architecture with convolutional, activation, pooling, and fully connected layers.
    - Configure the training process (loss function, optimizer).
3. **Model Training:**
    - Train the model using the preprocessed images.
    - Validate the model using a separate validation set to monitor performance and avoid overfitting.
4. **Model Evaluation:**
    - Test the trained model on unseen test data.
    - Evaluate model performance using metrics like accuracy, sensitivity, and specificity.
5. **Deployment:**
    - Deploy the model in a clinical setting for real-time analysis of lung images.

3.1 Pseudocode:(draft 1)

```
function preprocess_images(images):

    normalized_images = normalize(images)
    augmented_images = augment(normalized_images)
    return augmented_images


function build_cnn_model():

    model = initialize_model()
    model.add(ConvolutionalLayer(...))
    model.add(Activation('relu'))
    model.add(PoolingLayer(...))
    model.add(FullyConnectedLayer(...))
    model.compile(loss='categorical_crossentropy', optimizer='adam')
    return model


function train_model(model, training_data, validation_data):

    model.fit(training_data, epochs=50, validation_data=validation_data)
    return model


function evaluate_model(model, test_data):

    performance = model.evaluate(test_data)
    return performance


# Main execution flow

training_images, validation_images, test_images = load_dataset()
preprocessed_training = preprocess_images(training_images)
preprocessed_validation = preprocess_images(validation_images)
preprocessed_test = preprocess_images(test_images)

cnn_model = build_cnn_model()
trained_model = train_model(cnn_model, preprocessed_training, preprocessed_validation)
test_performance = evaluate_model(trained_model, preprocessed_test)
```

3.2 Pseudocode:(draft 2):

```
Initialize the software application

MainApplication:
  Display Welcome Screen
  Display Main Dashboard

Main Dashboard:
  Display options: Choose Dataset, Upload Image, Filter, Edit, Submit

Choose Dataset:
  Prompt user to select dataset for training the model
  Load selected dataset
  If dataset is valid:
    Display confirmation message
    Allow access to "Train Model" button
  Else:
    Display error message and prompt to choose again

Upload Image:
  Prompt user to upload lung image for analysis
  Retrieve image from user input
  If image format is valid:
    Display image with confirmation message
  Else:
    Display error message and request re-upload

Preprocess Image:
  Call ImageProcessor
  Normalize image
  Resize image to match model input requirements
  Convert image to grayscale (if required)
  Enhance image quality (contrast adjustment, noise reduction)
  Return preprocessed image

Train Model (if chosen dataset):
  Initialize CNN Model
  Split dataset into training and test sets
  Train CNN on training set
  Validate model accuracy on test set
  Update model parameters based on test results
  Save trained model
```

```
Analyze Image:
  If trained model exists:
    Load trained CNN model
    Input preprocessed image into CNN model
    Perform analysis to detect lung disease
    Generate diagnosis results
  Else:
    Display error message "Model not trained"

Generate Report:
  Collect analysis results
  Format results into a readable report
  Include details like diagnosis, confidence levels, and recommendations

Display Results:
  Show analysis results on GUI
  Provide option to view detailed report
  Allow user to save or print the report

End Program:
  Close application or return to Main Dashboard for new task
```

4. Expected Achievements

Lung Differentiation Accuracy: Achieve a high level of accuracy in distinguishing between the lungs of smokers and non-smokers. Set a specific target accuracy rate that reflects the reliability of the developed algorithm.

Disease Identification Precision: Develop a system capable of accurately identifying potential illnesses within the lungs, such as early signs of respiratory diseases associated with smoking. Define specific metrics for precision and recall to evaluate the system's diagnostic capabilities.

User-Friendly Interface: Create an intuitive and user-friendly interface for the software application, ensuring that healthcare professionals can easily interpret and utilize the results generated by the algorithm. User satisfaction and ease of integration into existing medical workflows should be key considerations. Scalability: Design the system to handle a diverse and large-scale dataset, demonstrating its scalability. This is crucial for real-world applications where the system may encounter a wide range of lung images.

Interpretability and Explainability: Incorporate features that enhance the interpretability and explainability of the algorithm's decisions. This is particularly important in the medical field, where understanding the reasoning behind diagnoses is crucial for gaining trust from healthcare practitioners.

Validation through Clinical Trials: If feasible, conduct clinical trials to validate the effectiveness of the developed software in a real-world medical setting. Obtain feedback from healthcare professionals and patients to refine and improve the system.

Privacy and Security Measures: Implement robust privacy and security measures to ensure the protection of sensitive medical data. Compliance with healthcare data protection regulations should be a priority.

5. Requirements

5.1 Functional Requirements (FRs)

1. **Image Processing Capability:** The software must be able to process lung images in formats such as PNG, JPEG, or DICOM.
2. **Disease Identification:** It should accurately identify specific lung diseases from processed images, including COPD, Emphysema, Lung Cancer, Bronchitis, and Pulmonary Fibrosis.
3. **Differentiation Function:** The system must differentiate between the lung images of smokers and non-smokers.
4. **Data Input Handling:** The application must accept input from various sources including direct uploads, electronic health records (EHR), and cloud storage.
5. **Reporting:** Generate detailed reports outlining the findings from the image analysis, including potential lung conditions and smoking status indications.
6. **User Interface Interactions:** Users should be able to interact with the system through a graphical user interface to upload images, view reports, and perform system configurations.

5.2 Non-Functional Requirements (NFRs)

1. **Usability:** The software interface should be intuitive and easy to use, requiring minimal training for healthcare professionals.
2. **Performance:** The image processing and analysis should return results within a few seconds to ensure efficiency in clinical workflows.
3. **Scalability:** The system must handle an increase in data volume from multiple healthcare facilities without degradation in performance.
4. **Reliability:** Achieve a high level of accuracy in disease detection and differentiation.
5. **Maintainability:** The software should be easy to update and maintain without significant downtime or disruptions.

6. Unique Features and Research/Engineering Challenges

Transfer Learning for Imaging Data: Explore and implement advanced transfer learning techniques tailored for medical imaging data. This could involve overcoming challenges related to domain adaptation and fine-tuning pre-trained models on a specific medical imaging dataset.

Handling Heterogeneous Data: Address the challenge of handling diverse lung images, considering variations in imaging modalities, resolutions, and data quality. Develop preprocessing techniques to standardize input data while preserving important diagnostic information.

Real-time Processing: Investigate and implement strategies for real-time processing of lung images, enabling quick and efficient analysis. This could involve optimizing algorithms and leveraging parallel processing capabilities.

Continual Learning for Dynamic Health Monitoring: Explore the integration of continual learning mechanisms to adapt the system over time. This is crucial for dynamic health monitoring, especially in cases where patients undergo changes in their smoking habits or develop new health conditions.

Ethical Considerations: Address ethical considerations related to the use of AI in healthcare, particularly in providing accurate information about smoking history. Develop strategies to mitigate potential biases and ensure fair and unbiased outcomes. Collaboration with Healthcare Professionals: Foster collaboration with healthcare professionals throughout the development process to align the software with practical clinical needs and ensure seamless integration into medical practices. Regular feedback loops should be established.

By achieving these milestones, the project aims to deliver a robust, accurate, and ethically sound software application that contributes significantly to the field of medical diagnostics related to smoking-associated lung damage.

7. Preliminary software engineering documentation

7.1 Engineering process

Development process till now: Until this time we have achieved a great milestone, we prepared our model, our software language, work schedule for coding, GUI's to know and imagine how our software will be and we prepared diagrams to understand how our software will work and take a look how to connect everything together.

Development process in future: our plan to future is to build dataset, writing codes and use models, train our software, prepare a form to test steps, test our software, fix things if needed, and for sure to finish the book for this final project phrase A. Also to set and think about the exact algorithms we will use in our project in phrase B.

Motivation: Our motivation for this approach is driven by the goal to create a robust, reliable solution that leverages the latest in machine learning for significant contributions to healthcare, particularly in the early detection and diagnosis of lung-related ailments. We believe that a CNN-based model is the most effective approach for this purpose, given its proven capabilities in image recognition and analysis.

7.2 Engineering product

Algorithms and models

In our software, we plan to implement a sophisticated algorithmic approach, primarily centered around Convolutional Neural Networks (CNNs)[1]. These are particularly adept at processing and analyzing complex image data, making them ideal for our lung image analysis. Additionally, we intend to explore the use of Support Vector Machines (SVMs)[3] for their classification strength, possibly integrating them with CNNs for enhanced performance. Complementary algorithms like Random Forests could be used for their ability to handle large datasets and K-Nearest Neighbors (KNN)[1][2] for their simplicity and efficacy in classification tasks. This blend of algorithms will allow us to comprehensively analyze lung images, ensuring high accuracy in detecting lung diseases and distinguishing between smokers and non-smokers. We will continuously evaluate and adapt these algorithms to optimize our software's diagnostic capabilities, keeping in line with the latest advancements in machine learning and artificial intelligence.

Data structure

Our data structure will be designed to efficiently handle and process lung image data for analysis. It will be organized to support the varying formats and resolutions of medical images. The structure will include metadata for each image, like patient demographics, smoking status, and any known lung conditions. This will facilitate effective training and testing of our machine learning models. Additionally, the data structure will be scalable to accommodate large datasets and robust to ensure data integrity and security, crucial for medical data handling. We'll also implement data preprocessing steps to enhance image quality for more accuracy.

Data Structure Overview:

1. **Image Data Storage:**
   - **Images:** Store the lung images in a standardized format (e.g., PNG or DICOM). Each image file should be named uniquely, ideally incorporating the patient ID and the date of the scan for easy tracking.
   - **Directory Structure:** Organize images in directories by patient ID and within those by date. This hierarchical structure facilitates easy access and management of patient data over time.
2. **Metadata Storage:**
   - **Patient Metadata:** This includes patient identifiers, demographic data (age, sex, etc.), and smoking history. Store this information in a secure, compliant database, linking each record to corresponding image data through patient IDs.
   - **Image Metadata:** Store details like image dimensions, acquisition parameters, and annotations (e.g., bounding boxes, labels of observed pathologies). This metadata is crucial for training and provides context for the model's predictions.
3. **Labels and Annotations:**
   - **Labels:** Each image should have associated labels indicating the presence or absence of specific lung conditions. These labels are used for training the CNN.
   - **Annotations:** Detailed annotations (e.g., locations of nodules or areas of fibrosis) made by radiologists can be stored in a structured format like JSON or XML. These annotations are invaluable for both training and validating the deep learning model.
4. **Data Split Structure:**
   - **Training, Validation, and Test Sets:** Organize the data into these three subsets to train and evaluate the CNN. Ensure that the split is representative of the overall dataset to avoid biased training outcomes.

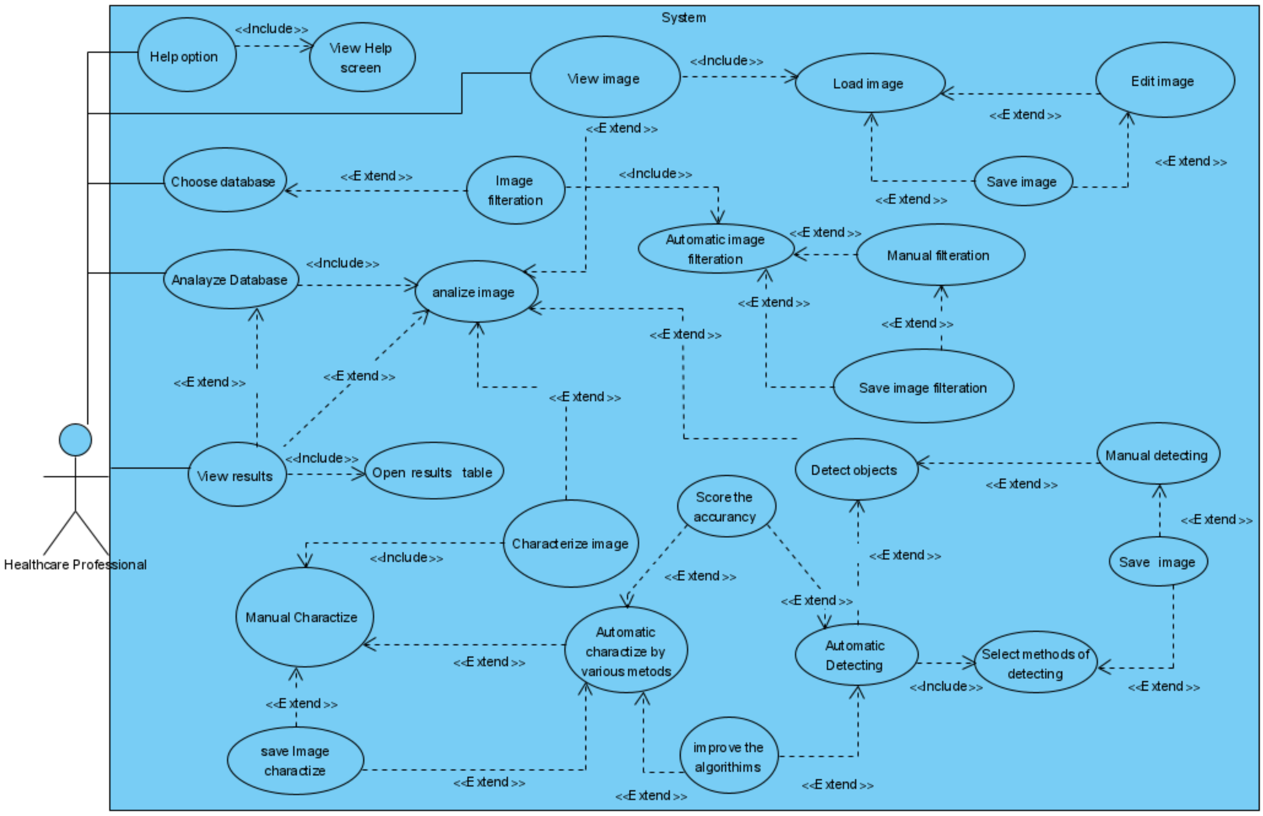
Implementation details

- **Data Format:** Use HDF5[23] or similar formats for efficient storage and access to large volumes of image data and metadata. These formats support large datasets and allow for rapid access and manipulation, which is ideal for deep learning applications.
- **Data Pipeline:** Implement a data pipeline using frameworks like TensorFlow[13] or PyTorch, which includes data loading, preprocessing (e.g., normalization, augmentation), and batching. This pipeline ensures that data is fed into the model in an efficient and effective manner during training and inference.

Why did we choose this structure?

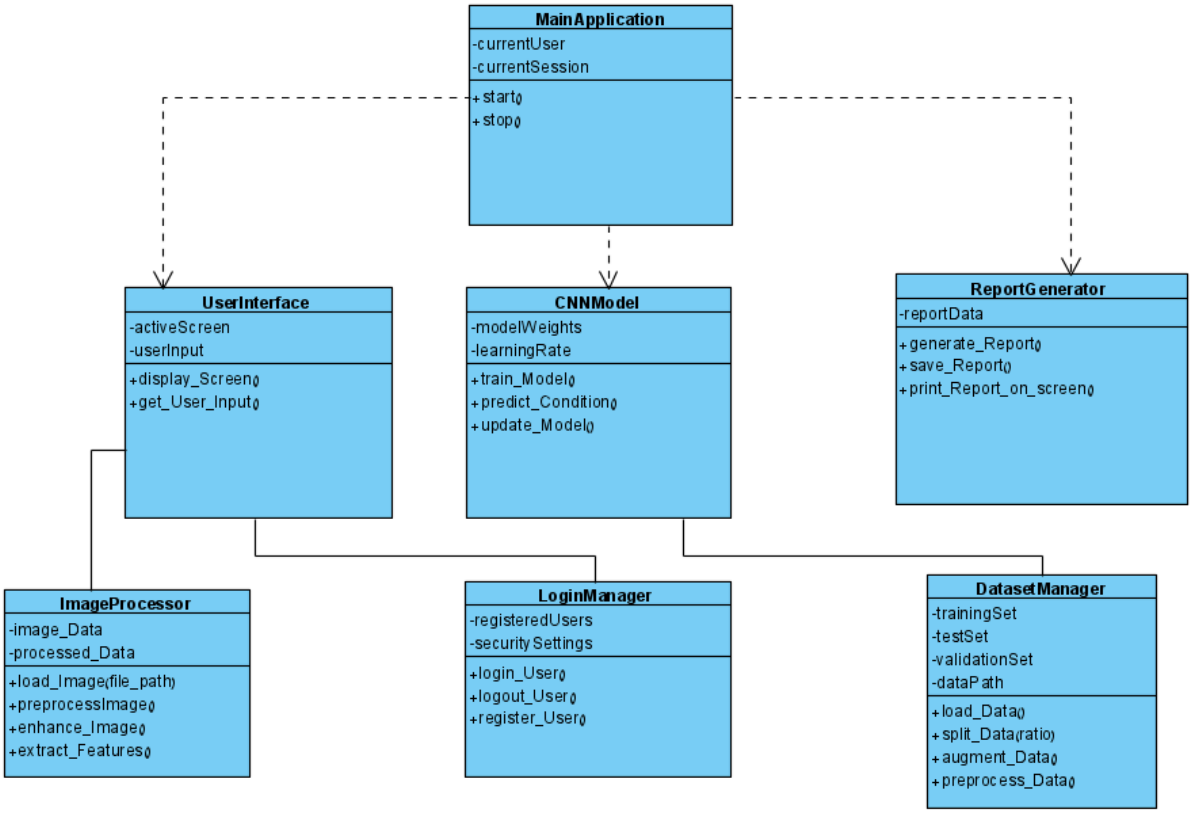
- **Scalability:** Easily scales to accommodate more patients, images, or detailed annotations.
- **Flexibility:** Adapts to various types of analyses, including both clinical assessments and deep learning applications.
- **Compliance:** Facilitates the management of sensitive health data in compliance with regulations like HIPAA.32

7.3 Use case diagram



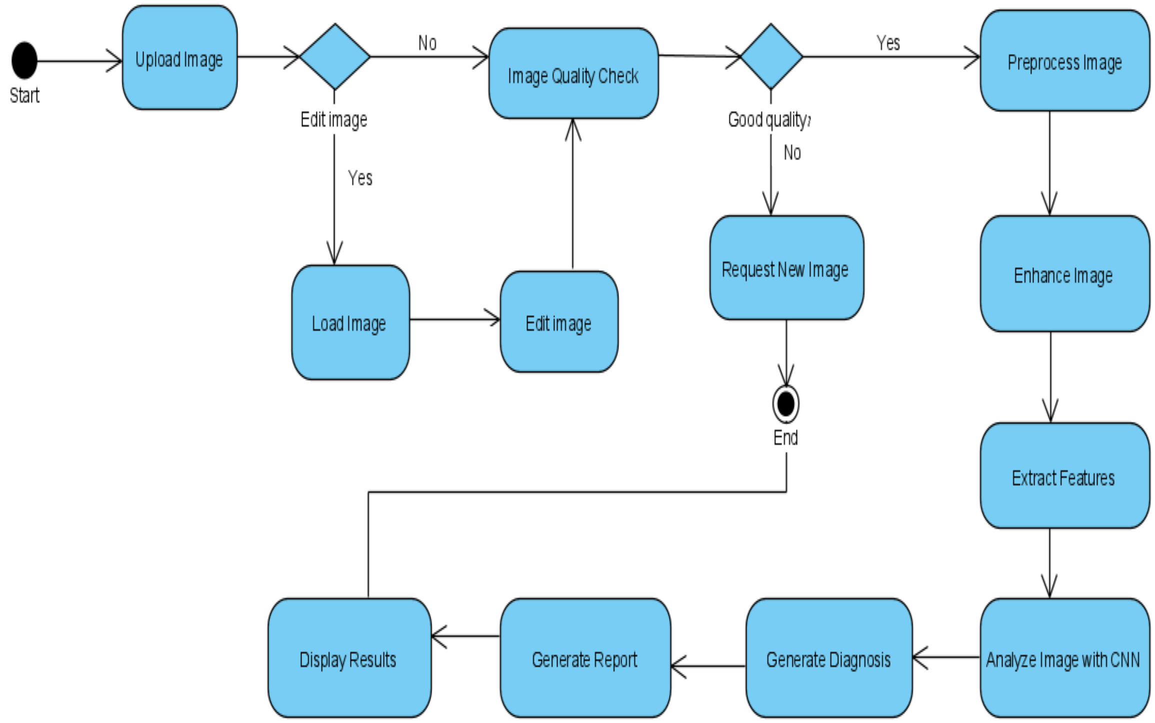
The image in better quality is available in https://ibb.co/m52ZfRf

7.4 Class diagram



The image in better quality is available in https://ibb.co/YjMFyQD

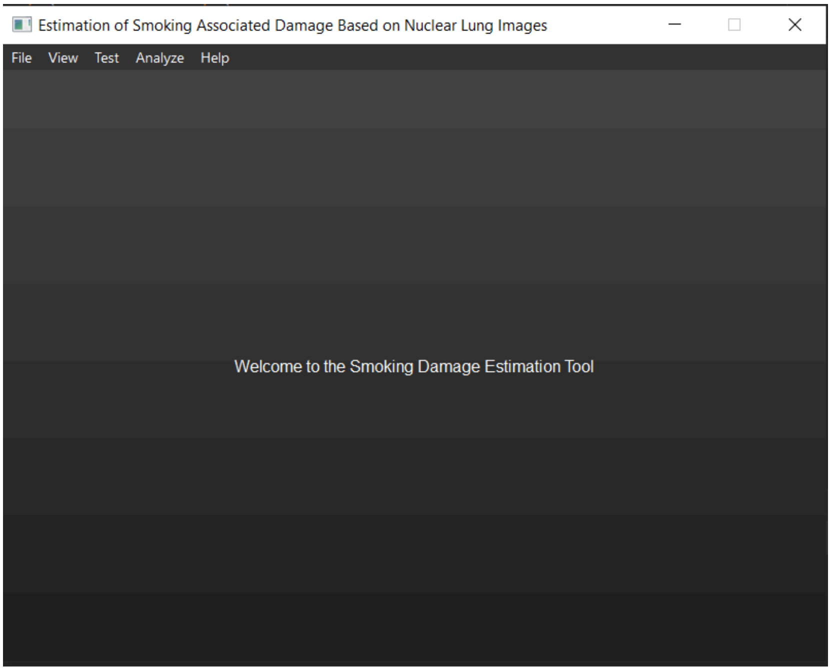
7.5 Activity diagram



The image in better quality is available in https://ibb.co/BBN3cVJ

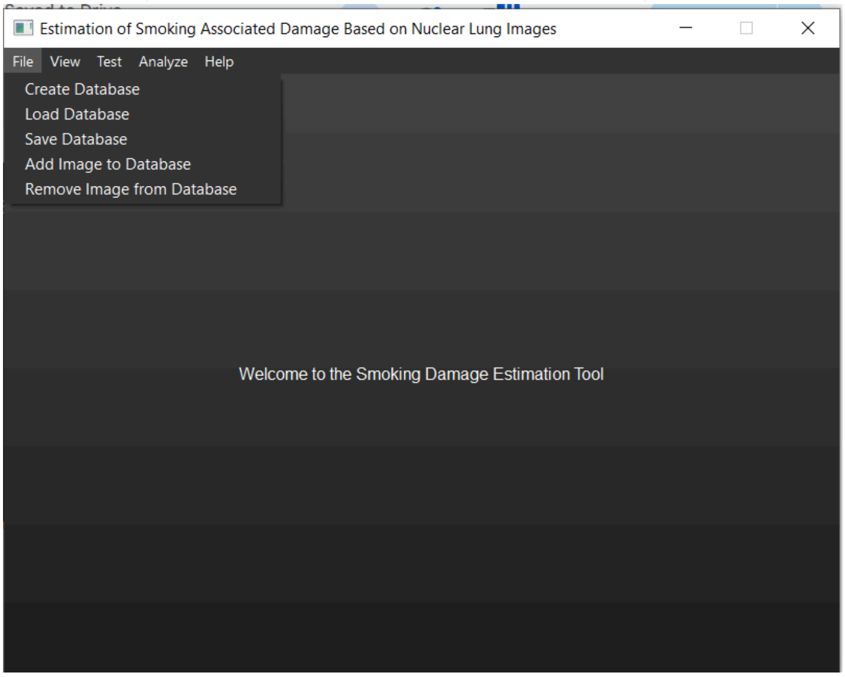
7.6 GUI

This is the main GUI when we run the code, we have the main bar at the top (e.g. file, view, test, analyze, and help).


Estimation of Smoking Associated Damage Based on Nuclear Lung Images

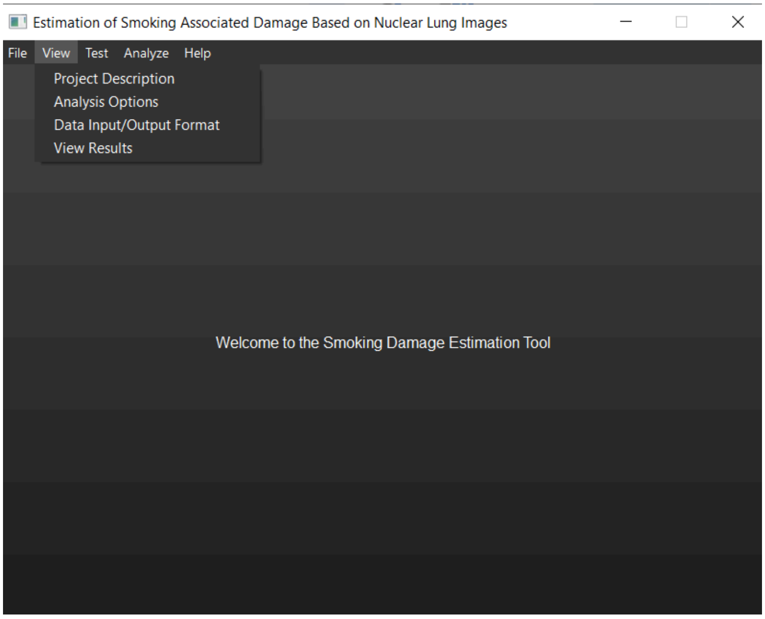
Welcome to the Smoking Damage Estimation Tool

This is the file option, we have 5 things we can do (e.g. create / load / save to the database, add an image to the database, and remove the image from the database).


Estimation of Smoking Associated Damage Based on Nuclear Lung Images

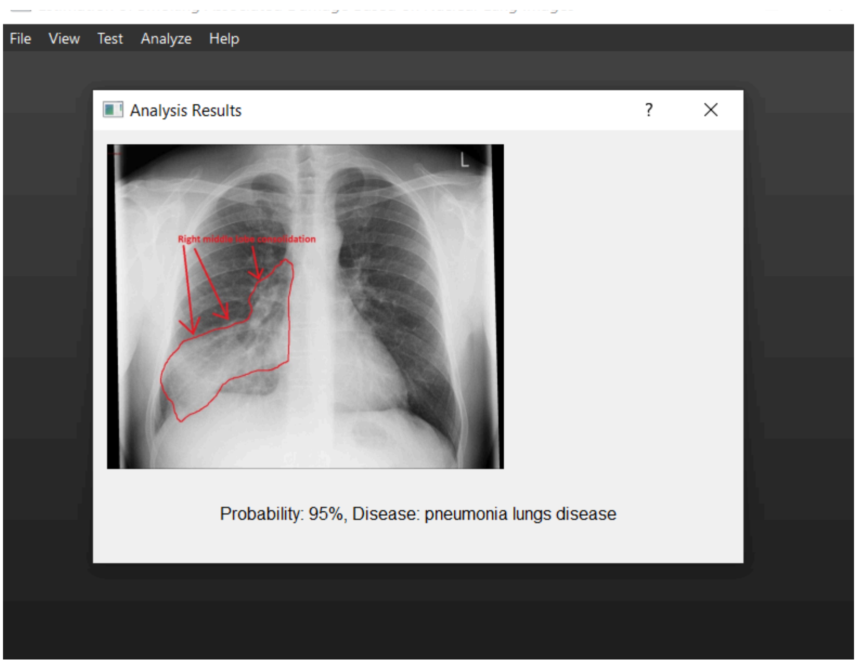
Welcome to the Smoking Damage Estimation Tool

This window shows the view options, we have project description, analysis options, and data input/output format.



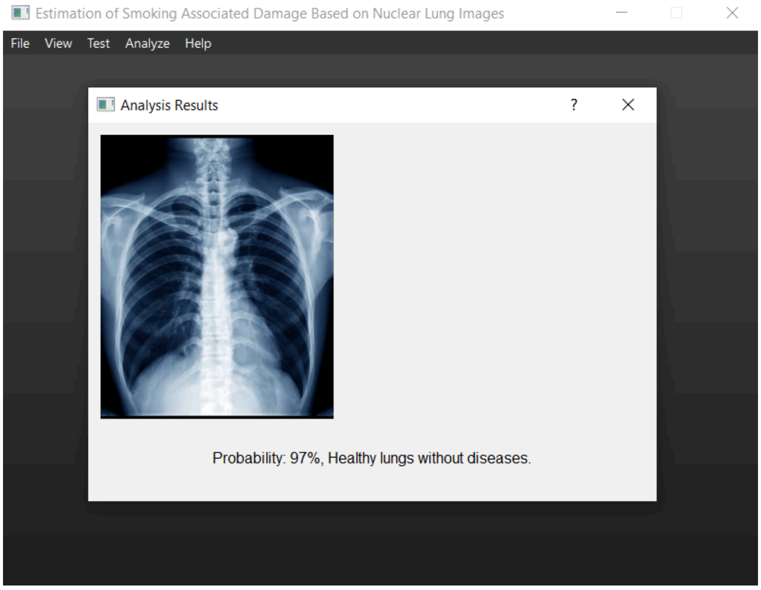
Welcome to the Smoking Damage Estimation Tool

This window shows the image of the lungs with the results (probability, disease name.) Here we have pneumonia lung disease.



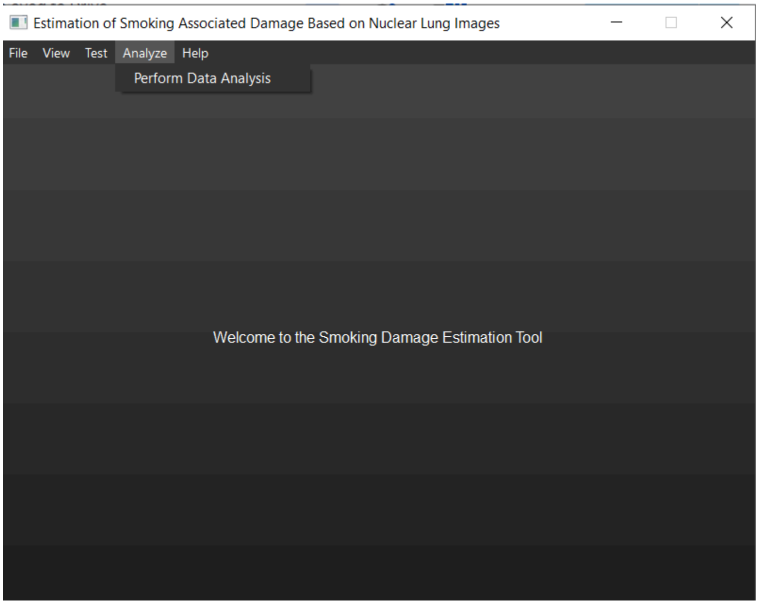
Probability: 95%, Disease: pneumonia lungs disease

Here in this window, we have healthy lungs, without disease:


Estimation of Smoking Associated Damage Based on Nuclear Lung Images
File View Test Analyze Help
Analysis Results

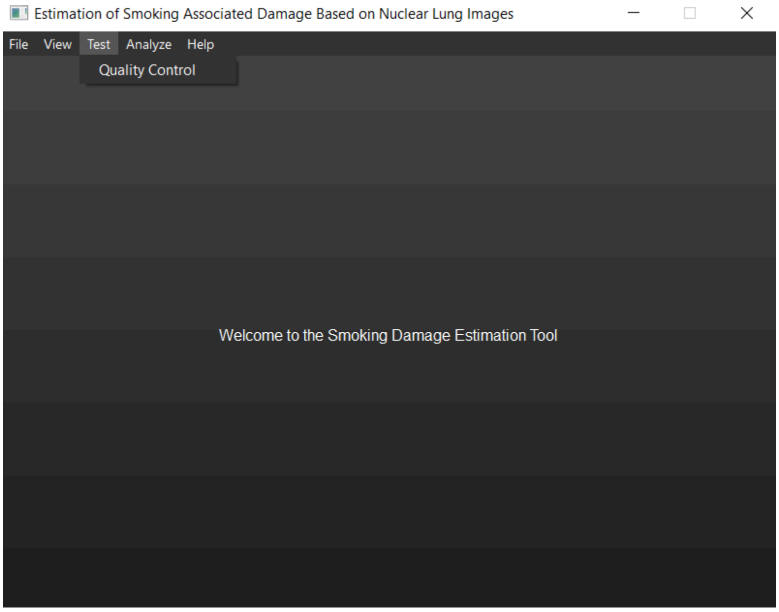
Probability: 97%, Healthy lungs without diseases.

This window shows the Analyze option, we have only performed data analysis in it.


Estimation of Smoking Associated Damage Based on Nuclear Lung Images
File View Test Analyze Help
Perform Data Analysis

Welcome to the Smoking Damage Estimation Tool

This window shows the test option, we have quality control in it.


Estimation of Smoking Associated Damage Based on Nuclear Lung Images
File
View
Test
Analyze
Help
Quality Control

Welcome to the Smoking Damage Estimation Tool

7.7 Testing Plan

1. **Objective:** Firstly we have to ensure that our application analyzes lung images, identifies anomalies, and differentiates between smokers and non-smokers with high precision and reliability.
2. **Test phases:**
   - **Unit Testing:** Individual components of the AI algorithms will be tested for functionality and performance.
   - **Integration Testing:** Test how integrated components (e.g., image preprocessing and AI analysis modules) function together.
   - **System Testing:** Evaluate the complete system's performance under various conditions.
   - **User Acceptance Testing (UAT):** Conducted with potential end-users to ensure the system meets user expectations and is usable in real-world scenarios.
3. **Data Collection:**
   - Dataset Acquisition: Secure diverse and comprehensive datasets including lung images from both smokers and non-smokers, with various lung conditions.
   - Data Annotation: Expert radiologists will annotate the images, identifying key features necessary for training the AI.
4. **Performance Metrics:**
   - **Accuracy:** Measure the percentage of correct diagnoses.
   - **Precision and Recall:** Evaluate the AI's ability to identify true positives and correctly ignore negatives.
   - **Robustness:** Test the AI's performance across different lighting, angles, and image qualities.
   - **User Feedback:** Gather qualitative assessments from radiologists regarding the AI's reliability and usability.
5. **Test Scenarios and Cases:**
   - **Positive Case Analysis:** Include images with clear signs of lung damage or smoking effects.
   - **Negative Case Analysis:** Include healthy lung images to test false positive rates.
   - **Edge Cases:** Test images that are not clear-cut, such as minor or early signs of disease.

References

1. Aggarwal, C. C. (2018). Neural Networks and Deep Learning: A Textbook. Springer.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
6. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.
7. Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)(n.d.). https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
8. National Institutes of Health (NIH). (2017). NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community. Retrieved from https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community
9. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252.
11. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
13. TensorFlow. (n.d.). TensorFlow: An end-to-end open-source platform for machine learning. Retrieved from https://www.tensorflow.org/
14. The Cancer Imaging Archive (TCIA). (n.d.). Retrieved from The Cancer Imaging Archive
15. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 3-11). Springer, Cham.
16. Lungs- Definition, Structure, Location, Functions, Diseases. Retrieved from https://microbenotes.com/lungs-structure-diseases/
17. Approach to the Chest x-ray (CXR). 5.2: Approach to the Chest x-ray (CXR) - Medicine LibreTexts.
18. Phenotypic comparison between smoking and non-smoking chronic obstructive pulmonary disease. Phenotypic comparison between smoking and non-smoking chronic obstructive pulmonary disease | Respiratory Research | Full Text
19. Smoking-related lung disease. Smoking-related lung disease | Radiology Reference Article | Radiopaedia.org

20. Chronic Obstructive Pulmonary Disease (COPD). Chronic Obstructive Pulmonary Disease (COPD)
21. Diffuse smoking-related lung diseases: insights from a radiologic-pathologic correlation. Diffuse smoking-related lung diseases: insights from a radiologic-pathologic correlation
22. Link to dataset: https://www.v7labs.com/open-datasets/chestx-ray14
23. Hierarchical Data Format. https://en.wikipedia.org/wiki/Hierarchical_Data_Format