

Capstone Project Phase A

Idiopathic pulmonary fibrosis classification using optimized Convolutional Neural Network

Alex Baboshin 310926415

Asaf Shnaider 316468636

(24-2-R-9)

Supervisor:

Prof. Miri Weiss Cohen

1. Introduction:

The term "usual interstitial pneumonia" (UIP) has evolved significantly since its introduction in the late 1960s. Initially, it described a wide range of acute and chronic lung diseases without a clear clinical correlation. Today, it is a specific histologic diagnosis closely associated with idiopathic pulmonary fibrosis (IPF) as Smith said [5].

Idiopathic Pulmonary Fibrosis (IPF) is a rare, chronic fibrosing interstitial pneumonia of unknown cause that significantly impacts both physical and emotional well-being. It is characterized by irreversible loss of lung function due to fibrosis, manifesting as a persistent cough, increasing dyspnea, and impaired quality of life, primarily affecting older adults​ as Maher said [1]​. Several factors may influence the risk of developing IPF, including age, family history, chronic gastrointestinal reflux, smoking, and certain environmental exposures. The average survival after diagnosis is 3–4 years, although some patients may experience a more rapid or slower progression as Hobbs said ​[2]​.

The American Thoracic Society (ATS) and European Respiratory Society (ERS) have previously issued guidelines on the diagnosis and management of IPF, which are now being reassessed. Main topics for reassessment include radiological and histopathological features of usual interstitial pneumonia (UIP), diagnostic criteria, and treatment approaches as Raghu said [3].The updated guidelines aim to define the progression of pulmonary fibrosis and determine the scope of antifibrotic therapy.

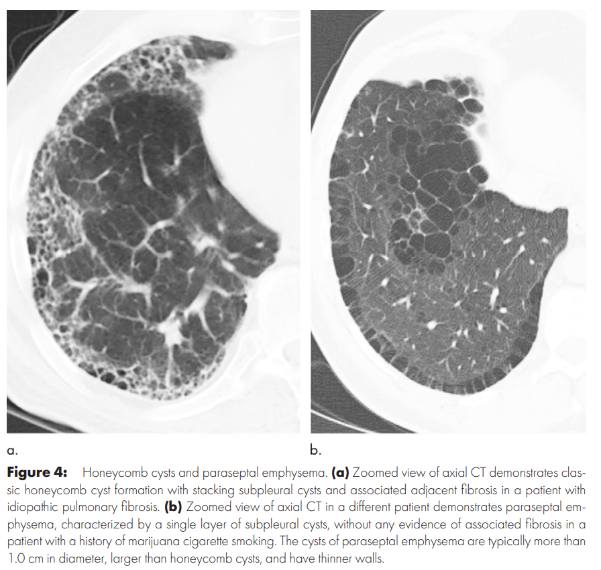
Pulmonary fibrosis can occur as a serious complication of viral pneumonia, such as from SARS-CoV or MERS-CoV infections, which can lead to long-term lung damage and impaired function. COVID-19, shares similarities with these viruses and has been associated with acute respiratory distress syndrome (ARDS) and potential fibrotic consequences. Whether COVID-19 can trigger irreversible pulmonary fibrosis deserves more investigation as Zou said [4]. Advanced imaging techniques, including AI-assisted systems, are being utilized to assess pulmonary inflammation and fibrosis progression, providing crucial data for clinical management and prognostic predictions as Zou said [4].

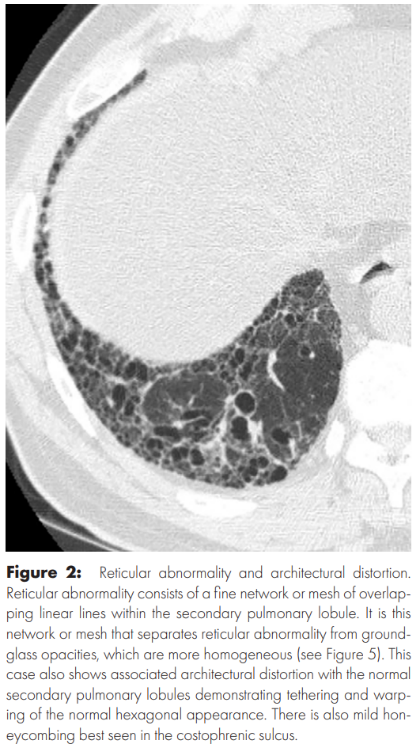
One article suggests key imaging findings necessary to evaluate a patient for IPF, highlighting few image definitions as significant indicators​ as Hobbs said [2]​.

Thin-section computed tomography (CT) has become a cornerstone in the evaluation of patients suspected of having IPF. This imaging technique provides high-resolution images of the lung, allowing for detailed visualization of interstitial abnormalities as Hobbs said [2]. Thin-section CT is particularly valuable in identifying characteristic patterns of fibrosis, such as honeycombing, reticular abnormalities, and traction bronchiectasis. These detailed images are crucial for differentiating IPF from other interstitial lung diseases and for assessing disease progression and response to therapy. The precision of thin-section CT enhances the accuracy of diagnosing IPF and helps guide clinical decisions regarding the use of antifibrotic medications (pirfenidone and nintedanib).

Serval key imaging definitions seen in patients with IPF such as Honeycomb cysts, which are subpleural clustered cystic air spaces. Although honeycombing might seem easy to identify, it can be challenging in practice, and different observers may disagree on its presence or absence as Hobbs said [2].

However, as noted in as Hobbs said [2] the categorization of UIP patterns, the presence of honeycombing is no longer considered a required feature in many cases of IPF. This alleviates some of the issues with the reliability of honeycombing identification in clinical practice.

**Figure 1:** Honeycomb cysts and paraseptal emphysema. (a) Zoomed view of axial CT demonstrates classic honeycomb cyst formation with stacking subpleural cysts and assosiated adjacent fibrosis in a patient with idiopathic pulmonary fibrosis. (b) zoomed view of axial CT in different patient demonstrates paraseptal emphysema ,characterized by a single layer of subpleural cysts.without any evidence of assosiated fibrosis in a patient with a history of marijuana cigarette smoking. The cysts of paraseptal emphysema are typically more than 1.0 cm in diameter,larger than honeycomb cysts, and have thinner walls.

****

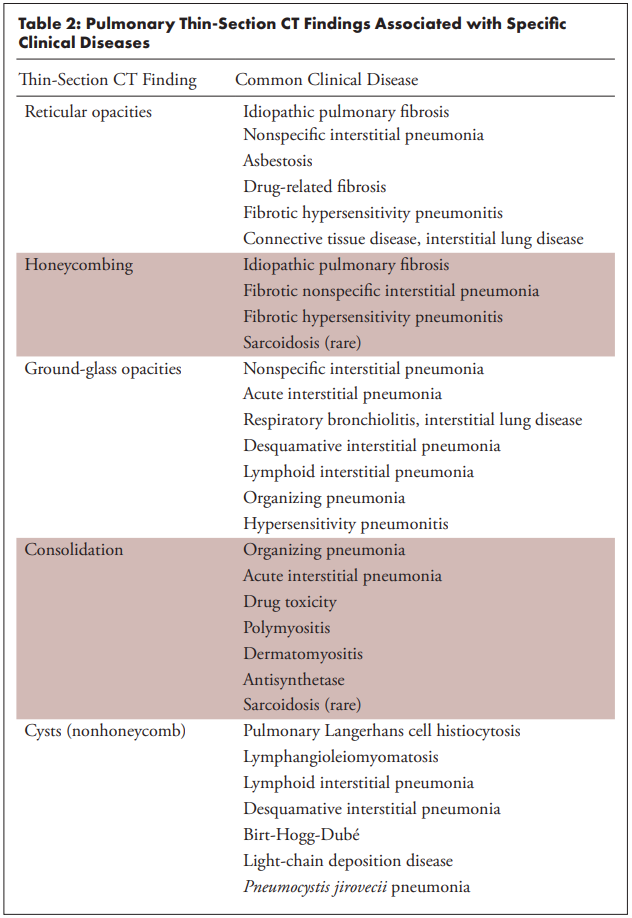
Another definition is Reticular pattern, also sometimes referred to as reticulation, consists of a fine network or mesh of overlapping linear lines within the secondary pulmonary lobule (Fig 2). Architectural distortion refers to any distortion of the normal lung parenchymal anatomy (Fig 2). In the setting of ILD, this generally refers to an abnormal appearance of the secondary pulmonary lobule shape or size with evidence of volume loss.

**Figure 2:** Reticular abnormality and architectural distortion. Reticular abnormality consists of a fine network or mesh of overlapping linear lines within the secondary pulmonary lobule. It is this network or mesh that separates reticular abnormality from groundglass opacities, which are more homogeneous (see Figure 5). This case also shows associated architectural distortion with the normal secondary pulmonary lobules demonstrating tethering and warping of the normal hexagonal appearance. There is also mild honeycombing best seen in the costophrenic sulcus.

The variety of different lung conditions in IPF patients and the similar symptoms in UIP patients suggests a difficult diagnosis of the disease. Accurate diagnosis guides patient selection for antifibrotic medications, which can reduce the decline in pulmonary function. This highlights the need for improved diagnostic techniques and optimization. Advancements in AI and machine learning hold promise for enhancing the accuracy and efficiency of IPF diagnosis, ultimately leading to better patient outcomes. Our research aims to explore these innovative technologies to develop more effective diagnostic tools for IPF.

The updated guidelines highlight the potential of computer-based quantitative CT (QCT) as a more objective and reproducible tool for assessing disease progression. QCT methods, including texture analysis and deep learning-based classification, have shown success in defining the extent of fibrosis and predicting mortality .

The integration of AI and convolutional neural networks (CNN) into diagnostic workflows offers significant enhancements in accuracy and efficiency. These advancements suggest that AI-assisted imaging not only aids in accurate diagnosis but also in monitoring disease progression and predicting patient outcomes .

1. Related Work:

Guidelines for Radiologic Diagnosis of IPF are suggested by Hobbs [2], Developed by the Radiology Working Group of the Pulmonary Fibrosis Foundation, these guidelines highlight the role of imaging in classifying UIP patterns, guiding biopsy decisions, and determining treatment options. Thin-section CT scans are essential for identifying characteristic fibrosis patterns, aiding in the accurate diagnosis of IPF as Hobbs said [2]. Researchers at National Jewish Health developed a process known as data-driven textural analysis, linking machine learning techniques with textural analysis to identify healthy and fibrotic lung at thin-section CT. The resulting data-driven textural analysis score can then be correlated to clinical risk factors, such as changes in pulmonary function. Rising data-driven textural analysis scores over time have been shown to be associated with increased disease progression and hospitalization as Hobbs said [2].

AI and machine learning techniques, such as convolutional neural networks (CNNs), enhance the extraction of quantitative data from biopsies, correlating histologic features with treatment outcomes. The integration of histologic criteria with molecular classifiers and advanced imaging techniques has improved the accuracy of UIP diagnosis, potentially reducing the need for surgical lung biopsies

The guidelines classify CT patterns into several categories. A typical UIP pattern shows subpleural and basal-predominant reticular abnormality with honeycomb cysts (figure 1). A probable UIP pattern is similar to typical UIP but without honeycombing. Indeterminate for UIP features are not sufficient for a firm UIP diagnosis. Non-IPF diagnosis patterns suggest other diseases, such as fibrotic hypersensitivity pneumonitis (HP), nonspecific interstitial pneumonia (NSIP), and connective tissue disease-related ILD (CTD-ILD). A multidisciplinary approach emphasizes the importance of collaboration among radiologists, pulmonologists, and pathologists in diagnosing and managing IPF.

The aim is to standardize the radiologic diagnosis of IPF, enhancing early and accurate detection through advanced imaging techniques. By integrating these practices into multidisciplinary discussions, the recommendations improve patient outcomes and guide appropriate treatment strategies. These guidelines provide a practical framework for radiologists, emphasizing the critical role of imaging in the accurate and early diagnosis of IPF and other fibrotic lung diseases. The structured approach and detailed imaging features help differentiate IPF from other conditions, ensuring better patient management and outcomes.

The diagnosis and classification of Idiopathic Pulmonary Fibrosis (IPF) have evolved significantly with advancements in imaging techniques and the integration of artificial intelligence (AI) and machine learning [3]. developed by a multidisciplinary committee using the GRADE approach. Updated criteria for IPF diagnosis emphasize key radiological features such as traction bronchiectasis and honeycombing seen in high-resolution computed tomography (HRCT) scans. Histopathological criteria confirm the diagnosis with features like patchy dense fibrosis and fibroblast foci. Transbronchial Lung Cryobiopsy (TBLC) is suggested as an acceptable, less invasive alternative to surgical lung biopsy (SLB) in centers with the appropriate expertise. With an 80% diagnostic yield compared to SLB’s 90%. TBLC is associated with fewer complications, such as pneumothorax and severe bleeding. The guidelines retain four HRCT categories for IPF diagnosis: UIP pattern, probable UIP pattern, indeterminate for UIP pattern, and alternative diagnosis. These categories help differentiate IPF from other fibrotic lung diseases.

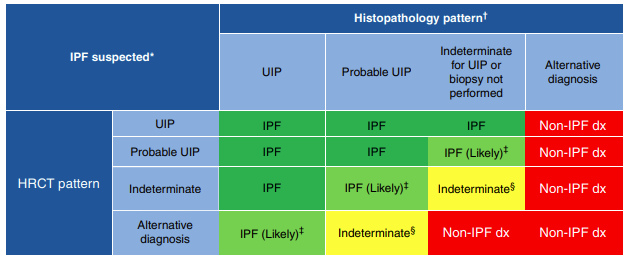


Figure 9. Idiopathic pulmonary fibrosis (IPF) diagnosis on the basis of high-resolution computed tomography (HRCT) and biopsy patterns, developed using consensus by discussion.

Raghu [3] provides several advancements in the diagnosis of IPF through machine learning and AI have been highlighted. The guidelines emphasize the evolving role of computer-based quantitative CT (QCT) in providing a more objective and reproducible measure of disease progression compared to visual assessment. The QCT techniques have advanced from simple histogram-based methods to more sophisticated machine learning approaches that include texture analysis, local histogram analysis, and deep learning-based classification. These methods have been successful in defining the extent and progression of the disease and predicting mortality​ as Zou said [4].

In addition to these imaging advancements, Forced Vital Capacity (FVC) remains a crucial physiological parameter in monitoring patients with IPF. It is most often used to follow disease progression because of its strong association with prognosis [3]. The guideline committee

recommends an absolute decline in FVC of more than 5% over one year as a key indicator of disease progression, a value that was extrapolated from the IPF literature. Although some trials have used a relative change in FVC to assess progression of pulmonary fibrosis, the committee prefers to use absolute change because it forecasts poorer outcomes and is regarded as an important predictor of mortality in IPF [3].

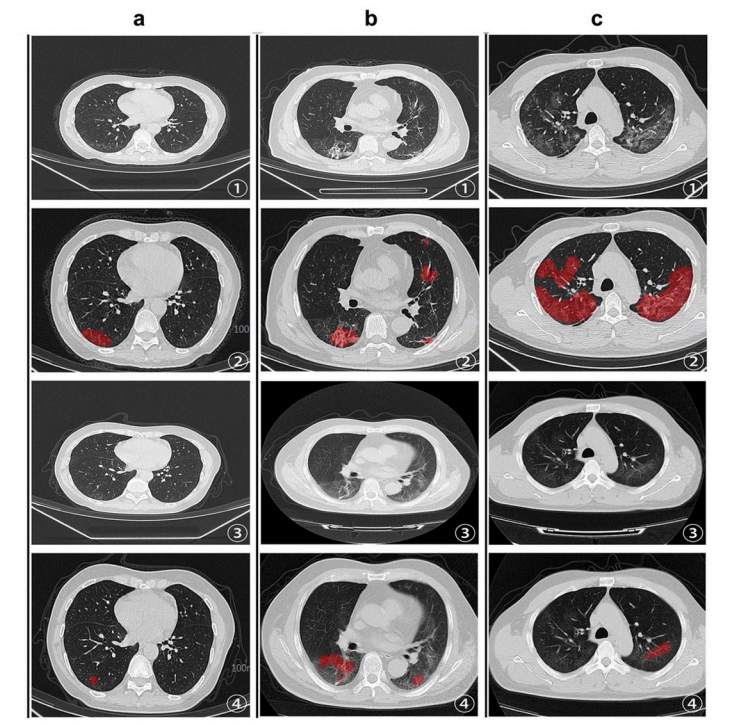
Deep learning techniques have shown promise in classifying fibrotic lung disease on high-resolution computed tomography (HRCT) scans, with studies demonstrating their ability to apply these methods to unsolved problems in imaging research​ as Zou said [4]. Furthermore, AI-assisted systems are described as objective tools that can qualitatively and quantitatively assess the progression of pulmonary inflammation​ as Zou said [4]. These advancements suggest that integrating AI and machine learning into the diagnostic process could significantly enhance the accuracy and predictive capabilities of current imaging techniques used in managing IPF.

Figure 10:Comparison of chest HRCT results between patients at discharge and 30 days after hospital discharge. a, b and c represent three separate patients: a represents a patient with mild pulmonary fibrosis, b represents a patient with moderate pulmonary fibrosis, and c represents a patient with severe pulmonary fibrosis.

Recent advancements in digital pathology and artificial intelligence (AI) have significantly enhanced the diagnostic capabilities in pulmonary pathology as Smith said [5]. AI tools are being developed to improve the accuracy of diagnosing idiopathic pulmonary fibrosis (IPF). For example, machine learning algorithms can analyze high-resolution computed tomography (HRCT) images and histopathological slides to identify patterns indicative of IPF. This capability highlights the potential of AI and machine learning to revolutionize the diagnosis and management of IPF by providing objective, reproducible assessments and uncovering prognostic biomarkers that may guide therapeutic decisions. The integration of AI in pathology not only augments the pathologist's interpretation but also allows the extraction of numerous quantitative data elements with diagnostic, prognostic, or therapeutic significance, adding significant value to histologic analysis in interstitial lung disease as Smith said [5].

The study [9] introduces a novel multi-resolution ConvNet framework designed for the automated segmentation of lungs with acute respiratory distress syndrome (ARDS). This framework leverages both high-resolution and low-resolution networks to balance local intensity and texture information with global contextual information.

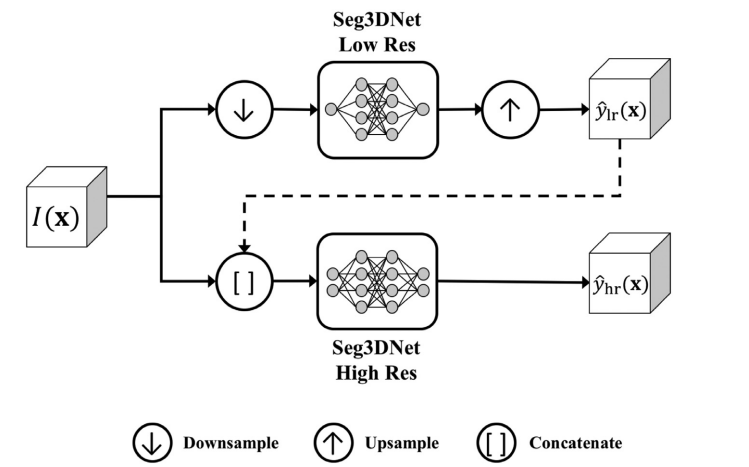
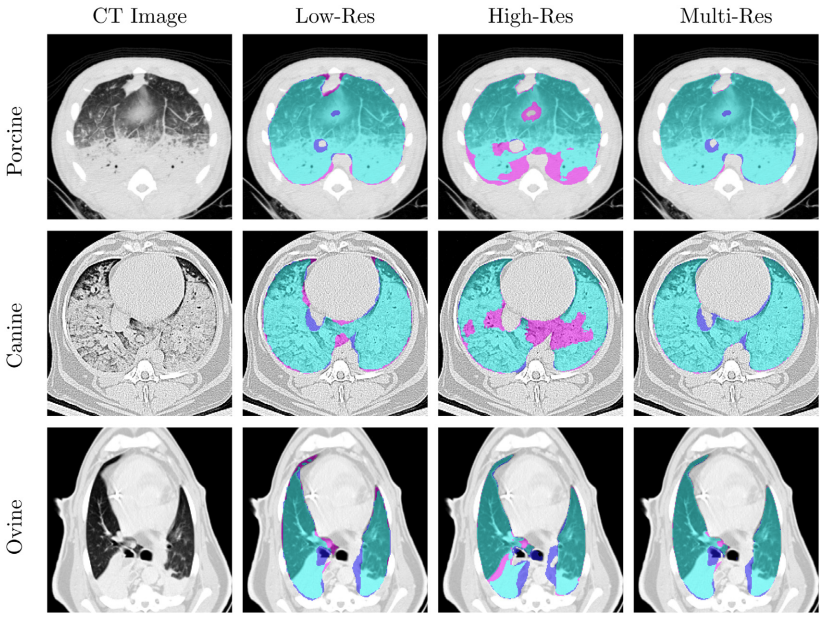
The underlying ConvNet architecture used in this study is a fully convolutional network (FCN) called Seg3Dnet. Seg3DNet, an extension of the U-Net architecture, is a fully convolutional network designed to handle three spatial dimensions. It consists of an encoder and decoder module that transform input images into abstract representations and reconstruct the segmentation map, respectively. The multi-resolution approach integrates low-resolution and high-resolution models, allowing the network to learn multi-scale features effectively.

Figure 11:Multi-resolution model. The upper pipeline corresponds to the low-resolution model and the lower pipeline corresponds to the high-resolution model. Global information learned in the low-resolution model was used in the high-resolution model, denoted by the dashed line.

Figure 12:Lung segmentation results for porcine, canine, and ovine subjects in the top, middle, and bottom rows, respectively. True positives, false negatives, and false positives are denoted in cyan, magenta, and purple, respectively.

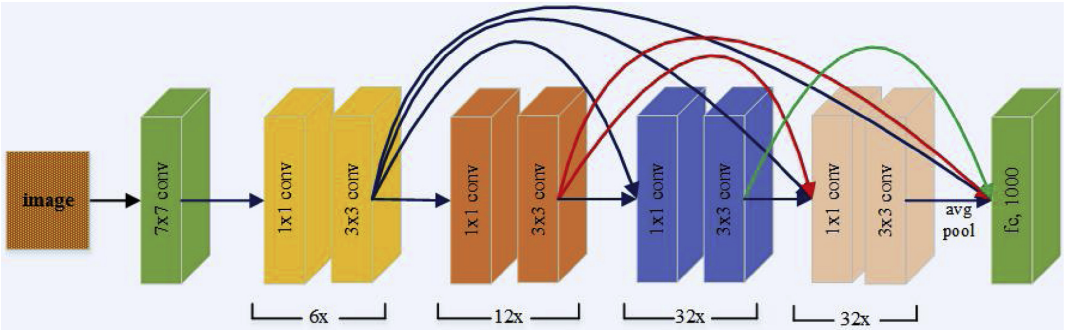
In this context, manual diagnosis of diseases from tissue samples by pathologists is time-consuming and expensive. Automated analysis of histopathology images can significantly improve early diagnosis and treatment, providing pathologists with a valuable second opinion [6].

The study of Talo [6] utilizes two pre-trained convolutional neural network (CNN) models, ResNet-50 and DenseNet-161, to classify histopathology images. These models were tested using both color and grayscale images.

1. Background

3.1 DenseNet 161

DenseNet-161, developed by Huang et al.[7], is characterized by its dense connectivity pattern, where each layer is connected to every other layer in a feed-forward manner. This design helps alleviate the vanishing gradient problem and encourages feature reuse, significantly reducing the number of parameters. DenseNet-161 achieved a classification accuracy of 97.89% on grayscale histopathology images, demonstrating its efficacy in medical image analysis. Its efficiency in training deep networks and reducing overfitting, especially with small datasets, makes it a powerful tool for histopathology image classification.

Figure 13:A block representation of DenseNet-161 Architecture [6].

ResNet-50, introduced by He et al.[8], incorporates skip connections that allow the network to jump over some layers. These residual connections help mitigate the vanishing gradient problem, enabling the training of very deep networks. ResNet-50 consists of 50 layers, making it a deeper network capable of learning more complex features from the data. It achieved the highest classification accuracy of 98.87% on color histopathology images, proving its robustness and efficiency. The architecture's ability to preserve learned features and improve convergence speed further enhances its performance in medical image classification tasks.

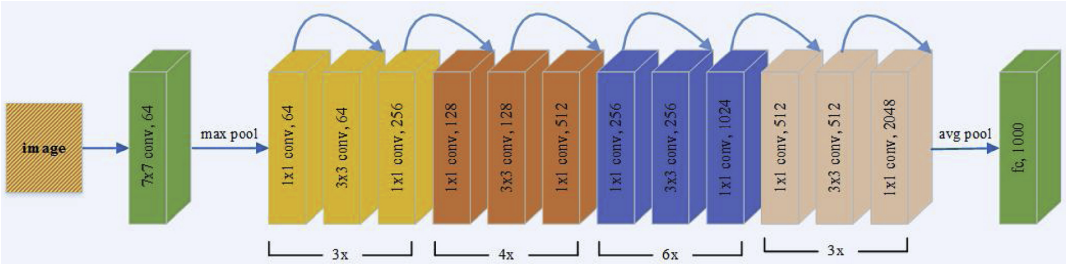


Figure 14:A block diagram representation of pre-trained Resneet-50 architecture [6].

As we noted DenseNet-161 achieved the best classification accuracy of 97.89% on grayscale images, while ResNet-50 achieved the highest classification accuracy of 98.87% on color images. These results indicate that the proposed pre-trained models can be used for fast and accurate classification of histopathology images, assisting pathologists in clinical tasks.

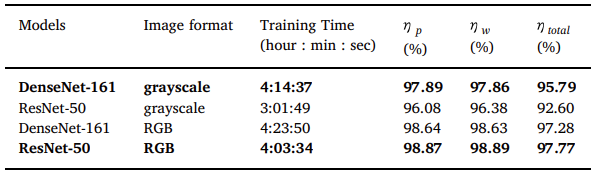


Figure 15:The results obtained using DenseNet-161 and ResNet-50 models on grayscale and color datasets [6].

The multi-resolution ConvNet framework, particularly the Seg3DNet architecture, can be adapted for the classification and segmentation of idiopathic pulmonary fibrosis (IPF) from CT images. By leveraging transfer learning and multi-scale feature learning, the framework can handle the complex and heterogeneous nature of IPF, providing accurate and efficient segmentation that aids in diagnosis and treatment planning.

However, The use of transfer learning and pre-trained CNN models like ResNet-50 and DenseNet-161 shows great promise for enhancing the classification of histopathology images, including those related to IPF. By incorporating these advanced techniques using AI and machine learning we can revolutionize the diagnosis and management of IPF.

3.2 Transfer learning:

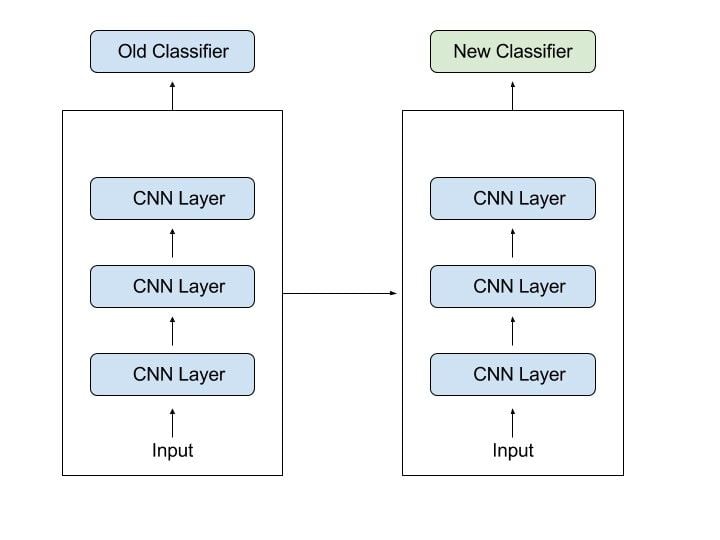
Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned [11]. While most machine learning algorithms are designed to address single tasks, the development of algorithms that facilitate transfer learning is a topic of ongoing interest in the machine-learning community. Transfer learning is described as "the reuse of a pre-trained model on a new problem." [12] It involves using the knowledge gained from one task to improve generalization in a related task. For example, a model trained to identify backpacks in images could transfer its learned features to recognize other objects like sunglasses. This approach is particularly valuable when there is limited data for the new task, as it allows leveraging the extensive knowledge already captured in a pre-trained model.

Figure 16: The left part of the diagram shows a pre-trained CNN model, which has been trained on a large dataset. The layers in this model have already learned features such as edges, textures, and shapes from the input data. the pre-trained CNN layers are retained and a new classifier is trained (right part of the diagram). The new classifier is fine-tuned on a smaller, domain-specific dataset.

1. Proposed Approach and research Process

4.1 Data set:

Data set used from kaggle OSIC Pulmonary Fibrosis Progression – Predict lung function decline [10], The data originally used to predict a patient’s severity of decline in lung function based on a CT scan of their lungs. Lung function is assessed based on output from spirometer, which measures the forced vital capacity (FVC), i.e. the volume of air exhaled.

In the dataset we are provided with a baseline chest CT scan and associated clinical information for a set of patients. A patient has an images acquired at time Week = 0 and has numerous follow up visits over the course of approximately 1-2 years, at which time their FVC is measured.

* In the training set, we are provided with an anonymized, baseline CT scan and the entire history of FVC measurements.
* In the test set, we are provided with a baseline CT scan and only the initial FVC measurement.

**Files**

The provided test set is a small representative set of files (copied from the training set) to demonstrate the format of the private test set.

* **train.csv** - the training set, contains full history of clinical information
* **test.csv** - the test set, contains only the baseline measurement
* **train/** - contains the training patients' baseline CT scan in DICOM format
* **test/** - contains the test patients' baseline CT scan in DICOM format

The original dataset is of file type DCM, we used custom python code to convert all files to PNG for ease of use.

As mentioned, A patient has an images acquired at time Week = 0 and has numerous follow up visits over the course of approximately 1-2 years

Patient with ID "ID00010637202177584971671" lung CT scan over the course of 54 weeks(1 year):

This is a 60-year-old male, an EX-smoker, who had scans every few weeks for 54 weeks

תמונה שמכילה שחור ולבן, מונוכרום, אומנות

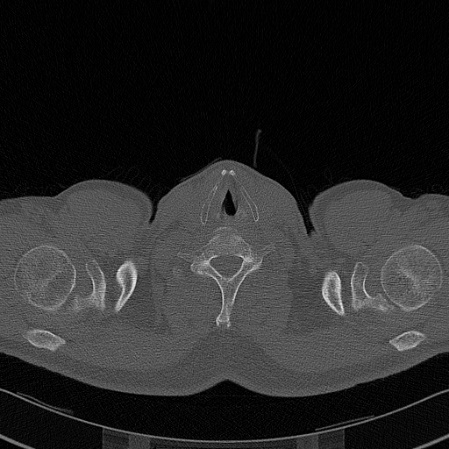
התיאור נוצר באופן אוטומטי ברמת מהימנות בינונית

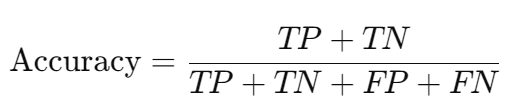
Figure 17: Patient’s CT scan, week = 0 (left) measures FVC = 3523 and percent = 94%. 6 months (middle) measures FVC = 2993 and percent = 80.47%. . Year = 1 (right) measures FVC = 2518 and percent = 67.7% of the typical FVC for a person of similar characteristics. The decline of lung capacity indicates a possible IPF in the patient.

* 1. Hyperparameter Optimization and Performance metrics

In the training phase, we  will examine the effect of hyper-parameters on model accuracy, in order to determine a minimum loss value.  To figure out which combination of these values works best, we'll test different combinations

* Learning rates will be tested at 5e-4, 5e-5, and 5e-6, and a loss function adaptation will be implemented.
* Epoch sizes range from 50 to 150.
* We will consider 32 vs. 64 batch size
* Dropout rate of 0.2 to 0.5.

 Based on sensitivity, specificity, precision, and recall, we will evaluate the performance of the classification system. The following measures are defined for evaluation:

1. **True Positives (TP)**: The number of correctly predicted positive cases (e.g., correctly identified IPF cases).
2. **False Positives (FP)**: The number of incorrect predictions where the model falsely identifies a negative case as positive (e.g., healthy lung identified as IPF).
3. **True Negatives (TN)**: The number of correctly predicted negative cases (e.g., correctly identified non-IPF cases).
4. **False Negatives (FN)**: The number of incorrect predictions where the model fails to identify a positive case (e.g., IPF case identified as healthy).
5. **Accuracy**: The ratio of correctly predicted instances (both positive and negative) to the total number of instances. It provides an overall measure of the model's correctness:
6. A mathematical equation with black text

   Description automatically generated**Precision**: The ratio of true positive predictions to the total predicted positives, indicating how many of the predicted positive cases were actually correct:
7. A mathematical equation with black text

   Description automatically generated**Recall (Sensitivity)**: The ratio of true positive predictions to the actual number of positive cases, showing how effectively the model identifies positive cases:
8. **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two, especially useful when dealing with imbalanced datasets:

A black text on a white background

Description automatically generated

1. Flow chart:

The process begins with uploading a training set consisting of CT images from various subjects or selecting pretrained model. The CNN is then trained using these images, where multiple convolutional layers extract features from the input data. Once the CNN is fully trained, the model is used to segment new test scans by identifying regions of interest (e.g., areas affected by IPF). After segmentation, the scans are classified to determine whether the patient is diagnosed as "Positive" or "Negative" for IPF.

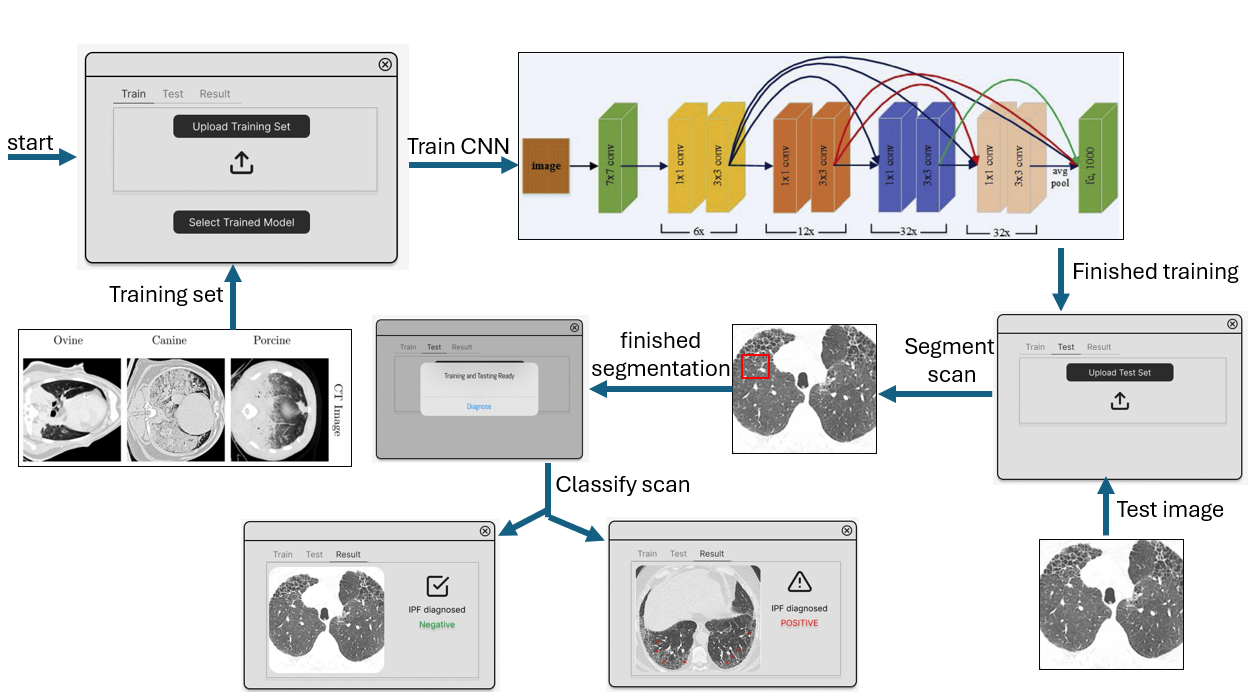


Figure 18: Flow chart process of the suggested model.

1. GUI:

A screenshot of a upload test set

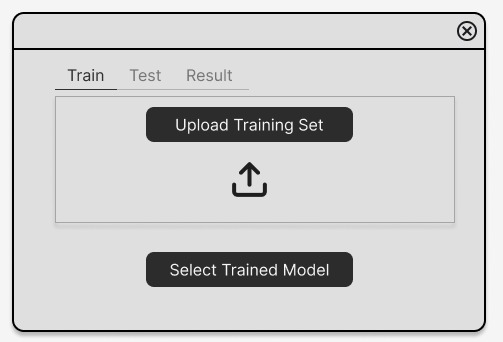
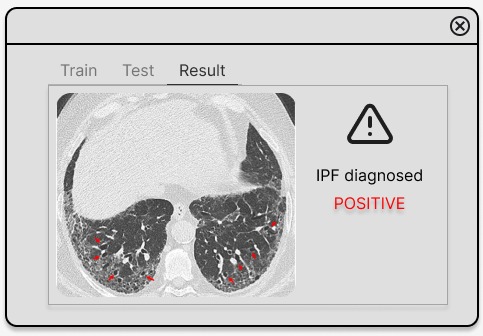
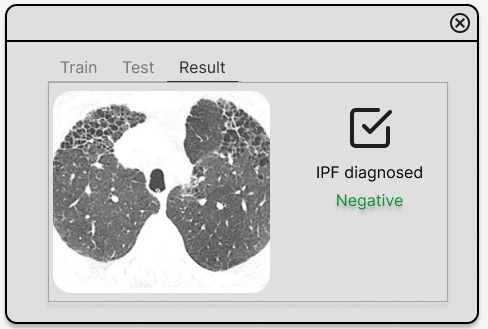
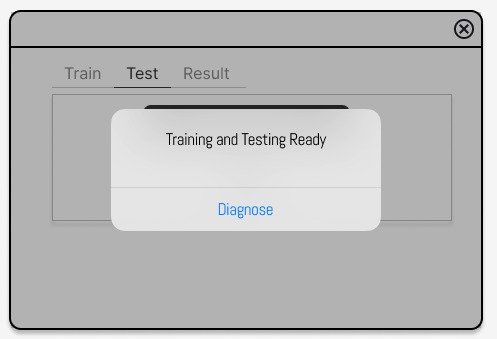
Description automatically generatedThe GUI consists of 3 tabs, Train tab for training the model, the user can either upload training set to be trained or select a previously trained model to be used. Second tab is Test, in this tab the user will upload the test set he wish to test and diagnose. Third tab is the result tab which will be the result diagnosed of the patient, either positive or negative IPF.

Figure 19: GUI images of suggested model to diagnose IPF in patients.

1. Evaluation / Verification Plan

|  |  |  |
| --- | --- | --- |
| Case # | Test explanation | Expected result |
| 1 | Insert wrong file type for "Upload Training Set" | Throw error: "Wrong file, please try again" |
| 2 | Insert expected file for "Upload Training Set" | The system will work with the training set we chose |
| 3 | Choose "Select Trained Model" | The system will work with the training set we chose |
| 4 | Insert wrong file type for "Upload Test Set" | Throw error: "Wrong file, please try again" |
| 5 | Insert expected file type for "Upload Test Set" | The system will get the scan and starts segment him |
| 6 | Press Diagnose | The system finished the segment and show us the IPF was diagnosed and show the scan |
| 7 | Press Diagnose | The system finished the segment and show us the IPF wasn't diagnosed and show us the scan |

References:

1. Maher, T. M., Bendstrup, E., Dron, L., Langley, J., Smith, G., Khalid, J. M., ... & Kreuter, M. (2021). Global incidence and prevalence of idiopathic pulmonary fibrosis. *Respiratory research*, *22*, 1-10.
2. Hobbs, S., Chung, J. H., Leb, J., Kaproth-Joslin, K., & Lynch, D. A. (2021). Practical imaging interpretation in patients suspected of having idiopathic pulmonary fibrosis: official recommendations from the Radiology Working Group of the Pulmonary Fibrosis Foundation. *Radiology: Cardiothoracic Imaging*, *3*(1), e200279.
3. Raghu, G., Remy-Jardin, M., Richeldi, L., Thomson, C. C., Inoue, Y., Johkoh, T., ... & Wilson, K. C. (2022). Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline. *American Journal of Respiratory and Critical Care Medicine*, *205*(9), e18-e47.
4. Zou, J. N., Sun, L., Wang, B. R., Zou, Y., Xu, S., Ding, Y. J., ... & Chen, S. M. (2021). The characteristics and evolution of pulmonary fibrosis in COVID-19 patients as assessed by AI-assisted chest HRCT. *PloS one*, *16*(3), e0248957.
5. Smith, M. L. (2022). The histologic diagnosis of usual interstitial pneumonia of idiopathic pulmonary fibrosis. Where we are and where we need to go. *Modern Pathology*, *35*, 8-14.
6. Talo, M. (2019). Automated classification of histopathology images using transfer learning. *Artificial intelligence in medicine*, *101*, 101743.
7. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
8. 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).
9. Gerard, S. E., Herrmann, J., Kaczka, D. W., Musch, G., Fernandez-Bustamante, A., & Reinhardt, J. M. (2020). Multi-resolution convolutional neural networks for fully automated segmentation of acutely injured lungs in multiple species. *Medical image analysis*, *60*, 101592.
10. <https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression/data>
11. Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI global.
12. https://builtin.com/data-science/transfer-learning