Department of Electrical & Computer Engineering

North South University

Senior Design Project

Prediction of Idiopathic Pulmonary Fibrosis Progression Using Deep Learning

---

Shazzad Hasan	**ID: 1530604043**
Md. Tanzim Hossain	**ID: 1620776042**
Md. Saidur Rahman	**ID: 1621529642**

Faculty Advisor

DR. TANZILUR RAHMAN

Assistant Professor

Department of ECE

Summer 2020

Table of Contents

Chapter 1: Introduction

In this chapter we will discuss about lung, what is idiopathic pulmonary fibrosis, what are the symptoms, what are the causes and risks, process of diagnosis, what are the treatments, what are the complications and at last we will talk about CT scan.

1.1 Lung

Lungs are sacks of tissue located just below the rib cage and above the diaphragm. They are an important part of the respiratory system and waste management for the body.

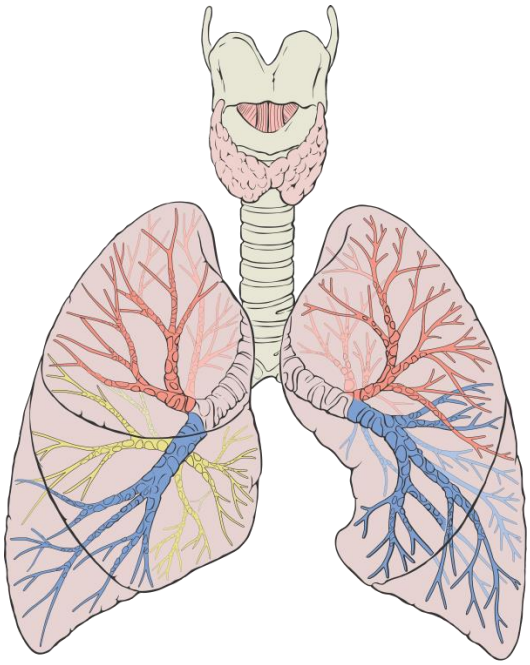

Figure 1: Lung

A person's lungs are not the same size. The right lung is a little wider than the left lung, but it is also shorter. The right lung is shorter because it has to make room for the liver, which is right beneath it. The left lung is narrower because it must make room for the heart.

Typically, a man's lungs can hold more air than a woman. At rest, a man's lungs can hold around 750 cubic centimeters of air, while a woman can hold around 285 to 393 cubic centimeters of air. The lungs are over-engineered to accomplish the job that we ask them to do. In healthy people without chronic lung disease, even at maximum exercise intensity, we only use 70 percent of the possible lung capacity.

The right lung is divided into three different sections, called lobes. The left lung has just two lobes. The lobes are made of sponge-like tissue that is surrounded by a membrane called pleura, which

separates the lungs from the chest wall. Each lung half has its own pleura sack. This is why, when one lung is punctured, the other can go on working.

As a person breathes, air travels down the throat and into the trachea, also known as the windpipe. The trachea divides into smaller passages called the bronchial tubes. The bronchial tubes go into each lung. The bronchial tubes branch out into smaller subdivisions throughout each side of the lung. The smallest branches are called bronchioles and each bronchiole has an air sac, also called alveoli. There are around 480 million alveoli in the human lungs. The alveoli have many capillary veins in their walls. Oxygen passes through the alveoli, into the capillaries and into the blood. It is carried to the heart and then pumped throughout the body to the tissues and organs. As oxygen is going into the bloodstream, carbon dioxide passes from the blood into the alveoli and then makes its journey out of the body. This process is called gas exchange. When a person breathes shallowly, carbon dioxide accumulates inside the body. This accumulation causes yawning.

The lungs have a special way to protect themselves. Cilia, which look like a coating of very small hairs, line the bronchial tubes. The cilia wave back and forth spreading mucus into the throat so that it can be dispelled by the body. Mucus cleans out the lungs and rids them of dust, germs and any other unwanted items that may end up in the lungs.

1.2 Idiopathic Pulmonary Fibrosis

Idiopathic Pulmonary Fibrosis (IPF) is a chronic, progressive and life-limiting condition. This condition causes scar tissue to build up in the lungs, which makes it more difficult for your lungs to work properly.

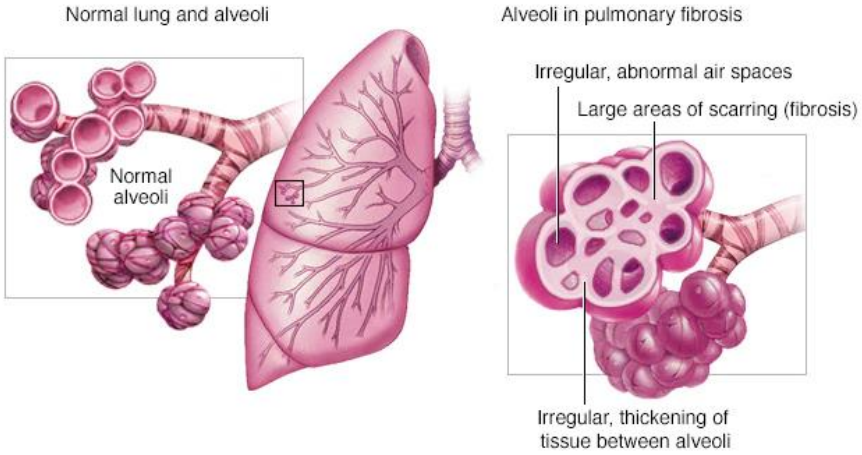

Figure 2: Lung with idiopathic pulmonary fibrosis

As pulmonary fibrosis worsens, you become progressively shorter of breath. In the end, IPF leads to life-threatening problems including respiratory failure. Progress rate can differ considerably

from person to person. In most cases, people experience respiratory problems, progressive scarring of the lungs, and a gradual reduction in lung function over the years. Quite frequently, for several years, infected patients have slight scarring in the lungs with little or no improvement in the condition. The condition can develop quickly (acutely) in some cases, causing life-threatening complications within a few years of diagnosis.

The word 'idiopathic' means unexplained or unproven the root cause of the condition, it has recently been shown that genetic vulnerability accounts for 35-40 percent of the risk of developing this disease. While there is no cure for IPF, there are several therapies available to control the condition and a range of potential therapeutic approaches are being explored. Ultimately, a lung transplant would require certain affected people.

1.2.1 Symptoms of Idiopathic Pulmonary Fibrosis

No symptoms may occur in the early stages of IPF. The development of the condition as mentioned above is highly variable. Some people may experience 'exacerbation' in which symptoms intensify for a period of time, before some improvement. The original, characteristic symptom is shortness of breath that is particularly noticeable during exercise. This is known as dyspepsia, or breathlessness. A moderate, dry cough that produces little or no sputum (non-productive cough) can also be seen by affected persons. For more than 30 days, this constant, non-productive cough persists.

As the condition progresses, upon significant exertion or exercise, affected individuals experience breathlessness. They may show rapid, shallow breathing. There may also be the dry, coughing, non-productive cough. Breathlessness will probably grow upon minimal exertion or even at rest. Individuals affected can experience repeated, unmanageable bursts of coughing.

Additional symptoms that may arise include abnormal fatigue, chest pain, slow unintended weight loss and painful joints and muscles. Some people can develop clubbing of the toes or fingers. Clubbing is when the tissue swells at the bottom of the fingernails and toenails, becoming broader and more oval. Individuals affected have an increased risk of repeated chest infections (chronic pneumonia).

Ultimately, respiratory function in individuals with IPF declines to cause severe complications including respiratory failure. Pulmonary fibrosis can lead to other severe medical conditions including pneumonia (lung infection), collapsed lungs (pneumothorax), high blood pressure of the main artery of the lungs (pulmonary hypertension), blood clots in the lungs (pulmonary embolism), and heart failure. Individuals with IPF may be at an increased risk of developing lung cancer.

Some individuals experience an ‘acute exacerbation,’ which describes a rapid progression of the disease and a rapid deterioration of lung function. Acute exacerbations may be associated with a complicating factor such as an infection, pulmonary embolism, pneumothorax or heart failure. However, in many cases, acute exacerbations occur without any identifiable cause.

1.2.2 Causes and Risks of Idiopathic Pulmonary Fibrosis

There is no complete understanding of the precise, underlying cause of IPF. The disease exists in households, often sporadically as well. Various factors, including immunological, environmental, and genetic factors, are thought to play a role in the disorder 's development. A mutation in the MUC5B gene is the main risk factor representing 30 per cent of the risk of developing IPF. MUC5B gene encodes a member of the mucin family of proteins, which are highly glycosylated macromolecular components of mucus secretions. This family member is the major gel-forming mucin in mucus. That results in more mucus production in the smallest airways in the lung (respiratory bronchioles).

Researchers have assumed for several years that most cases arise from widespread inflammation in the lungs, which has evolved to cause excessive scarring in the lungs. Researchers now conclude, however, that most cases result from damage to some cells lining the tiny airways and alveoli (epithelial cells). The alveoli are small, thin-walled air sacs that are located in the lungs in massive quantities. Alveoli are where the blood flows oxygen, and the blood releases carbon dioxide. At the ends of short, narrow tubes called bronchioles, which branch off from the main airway passages within the lungs. Basically, air is breathed in through the nose and mouth and travels down the throat to the windpipe (trachea). The trachea divides into air passages called bronchial tubes to which the bronchioles are connected. Most likely, as a part of normal wound healing, the body attempts to repair the damaged epithelial cells. This response is abnormal leading to progressive scarring and damage to the alveoli and surrounding lung tissue.

As explained, the underlying reason why the initial harm happens isn't continuously understood. Such harm could result from chronic exposure to Associate in Nursing inciting or ‘triggering’ agent. cigarette smoking is powerfully related to IPF, notably in people with a minimum of twenty ‘pack’ years of smoking history. extra triggering agents embrace chronic inhaling to the lungs of foreign material (chronic aspiration) and therefore the chronic respiratory in of sure environmental pollutants together with numerous gases and fumes, inorganic dusts (e.g. silicon dioxide and laborious metal dusts), and organic dusts (e.g. microorganism and animal proteins). infective agent or microorganism infections, radiation therapies, and sure medications together with specific chemotherapeutical medication, antibiotics and heart medications have additionally been connected to IPF. reaction diseases like rheumatism, lupus or scleroderma square measure acknowledged to be related to pneumonic pathology. In several cases, no inciting or triggering agent are often known.

In 5-10% of the cases, IPF has occurred in additional than one member of the same family unit (i.e. parent, kids and siblings). once this happens, the term familial upset pneumonic pathology is employed. The symptoms and objective signs of familial IPF square measure an equivalent as those for infrequent IPF, however the disorder tends to occur at a rather younger age.

Factors that make someone more susceptible to pulmonary fibrosis include:

- **Age:** Although pulmonary fibrosis has been diagnosed in children and infants, the disorder is much more likely to affect middle-aged and older adults.
- **Sex:** Idiopathic pulmonary fibrosis is more likely to affect men than women.
- **Smoking:** Far more smokers and former smokers develop pulmonary fibrosis than do people who have never smoked.
- **Certain occupations:** You have an increased risk of developing pulmonary fibrosis if you work in mining, farming or construction or if you're exposed to pollutants known to damage your lungs.
- **Cancer treatments:** Having radiation treatments to your chest or using certain chemotherapy drugs can increase your risk of pulmonary fibrosis.
- **Genetic factors:** Some types of pulmonary fibrosis run in families, and genetic factors may be a component.

1.2.3 Diagnosis of Idiopathic Pulmonary Fibrosis

A diagnosis of idiopathic pulmonary fibrosis can be assumed based on identifying the characteristic symptoms, a clear history of the patient, and a comprehensive clinical examination. A diagnosis can be confirmed based on a number of specialized examinations, including conventional chest X-rays (radiography), CT scans, Pulmonary function tests, blood tests, and surgical removal and microscopic lung tissue analysis (lung biopsy).

1.2.4 Treatments of Idiopathic Pulmonary Fibrosis

Traditional x-rays of the chest can reveal scarring in the lungs which is indicative but not IPF diagnosis. At the time of diagnosis, certain people may have regular chest x-rays. To diagnose individuals with IPF, a special form of CT scanning known as high resolution computed tomography (HRCT) may be used. A device and x-rays are used during CT scanning to produce a film that displays cross-sectional images of some tissue structures. HRCT offers clearer, more accurate photographs of the lungs than traditional x-rays or standard CT scans. The presence of scar tissue and the degree of lung damage can be revealed by HRCT, and in certain cases the presence of clear results can be sufficient to determine a diagnosis. Many IPF cases include a distinct pattern of lung damage known as typical interstitial pneumonia (UIP). This pattern consists of patches of natural lung tissues, which contrast with thick scar tissue patches (fibrosis).

Pulmonary function tests may also be useful to measure how efficiently the lungs absorb and exhale oxygen, and how easily they transfer oxygen to the blood. There are no IPF blood tests, but other factors may help to rule out such blood tests. Exercise monitoring that measures blood pressure, levels of oxygen saturation and heart function may be recommended.

A procedure called bronchoalveolar lavage (BAL) can help to rule out other conditions. A narrow tube (bronchoscope) is slipped down the windpipe into the lungs during BAL, and a sterile solution is passed through the tube that washes out cells. This fluid is collected and the tube is then removed

enabling examination of the cells. If further testing cannot confirm a diagnosis of IPF, a lung biopsy or a video-assisted thoracoscopy may be needed. A lung biopsy requires the removal from many locations inside the lungs of samples of lung tissue. A lung biopsy will rule out particular conditions and confirm an IPF diagnosis.

Video-assisted thoracoscopy involves placing, through a very small cut (incision) in the chest wall, a narrow tube called an endoscope attached to a small camera. This allows physician to examine the lungs or other structure within the chest cavity.

- **Radiation treatments**

Some people who receive radiation therapy for lung or breast cancer show signs of lung damage months or sometimes years after the initial treatment. The severity of the damage may depend on:

- How much of the lung was exposed to radiation.
- The total amount of radiation administered.
- Whether chemotherapy also was used.
- The presence of underlying lung disease.

- **Medications**

Many drugs can damage your lungs, especially medications such as:

- **Chemotherapy drugs:** Drugs designed to kill cancer cells, such as methotrexate (Trexall, Otrexup, others) and cyclophosphamide, can also damage lung tissue.
- **Heart medications:** Some drugs used to treat irregular heartbeats, such as amiodarone (Cordarone, Nexterone, Pacerone), may harm lung tissue.
- **Some antibiotics:** Antibiotics such as nitrofurantoin (Macrobid, Macrodantin, others) or ethambutol can cause lung damage.
- **Anti-inflammatory drugs:** Certain anti-inflammatory drugs such as rituximab (Rituxan) or sulfasalazine (Azulfidine) can cause lung damage.

1.2.5 Complications of Idiopathic Pulmonary Fibrosis

Complications of pulmonary fibrosis may include:

- **High blood pressure in your lungs (pulmonary hypertension).** Unlike systemic high blood pressure, this condition affects only the arteries in your lungs. It begins when the smallest arteries and capillaries are compressed by scar tissue, causing increased resistance to blood flow in your lungs. This in turn raises pressure within the pulmonary arteries and the lower right heart chamber (right ventricle). Some forms of pulmonary

hypertension are serious illnesses that become progressively worse and are sometimes fatal.

- **Right-sided heart failure:** This serious condition occurs when your heart's lower right chamber (ventricle) has to pump harder than usual to move blood through partially blocked pulmonary arteries.
- **Respiratory failure:** This is often the last stage of chronic lung disease. It occurs when blood oxygen levels fall dangerously low.
- **Lung cancer:** Long-standing pulmonary fibrosis also increases your risk of developing lung cancer.
- **Lung complications:** As pulmonary fibrosis progresses, it may lead to complications such as blood clots in the lungs, a collapsed lung or lung infections.

1.3 CT scan

A computerized tomography (CT) scan combines a series of X-ray images taken from different angles around your body and uses computer processing to create cross-sectional images (slices) of the bones, blood vessels and soft tissues inside your body. CT scan images provide more-detailed information than plain X-rays does.

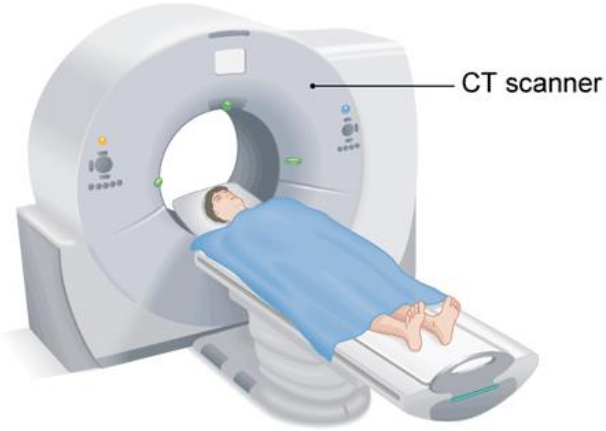

Figure 3: CT scan

A CT scan has many uses, but it’s particularly well-suited for diagnosing diseases and evaluating injuries. The imaging technique can help doctor to:

- diagnose infections, muscle disorders, and bone fractures.
- diagnose interstitial lung disease including idiopathic pulmonary fibrosis.
- pinpoint the location of masses and tumors (including cancer).
- study the blood vessels and other internal structures.
- assess the extent of internal injuries and internal bleeding.
- guide procedures, such as surgeries and biopsies.
- monitor the effectiveness of treatments for certain medical conditions, including cancer and heart disease.

1.4 Motivation

Imagine one day, your breathing became consistently labored and shallow. Months later you were finally diagnosed with pulmonary fibrosis, a progressive disease that naturally gets worse over time with no known cause and no known cure, created by scarring of the lungs. If that happened to you, you would want to know your prognosis. That's where a troubling disease becomes frightening for the patient. Outcomes can range from long-term stability to rapid deterioration, Natural history of IPF is unknown and the prediction of disease progression at the time of diagnosis is notoriously difficult and doctors aren't easily able to tell where an individual may fall on that spectrum. Data science, may be able to aid in this prediction. If successful, patients and their families would better understand their prognosis when they are first diagnosed with this incurable lung disease. Improved severity detection would also positively impact treatment trial design and accelerate the clinical development of novel treatments.

1.5 Aim and Objective

Lung function is assessed based on output from a spirometer, which measures the forced vital capacity (FVC), i.e. the volume of air exhaled. Our aim is to predict a patient's severity of decline in lung function based on a CT scan of their lungs, metadata, and baseline FVC as input. We want to predict the final three FVC measurements for each patient, as well as a confidence value in our prediction.

- **Patient Week:** a unique Id formed by concatenating the Patient and Weeks columns (i.e. ABC_22 is a prediction for patient ABC at week 22).
- **FVC:** the predicted FVC in ml.
- **Confidence:** a confidence value of your prediction (also has units of ml).

Chapter 2: Literature Review

For conducting our research project, we have explored a lot of paper that are related and not related to our work. We have studied them in order to explore how these research papers handle those challenges that we are facing now. While we were exploring, we have found a few papers which conducted deep learning approach on medical data specially on CT scan images. We chose some particular papers because their working approach is closely related to our work. We also deduced some ideas from these papers for conducting our work.

A paper titled "Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization - Random forest approach" [1] by Yu Shi was published in 2019 which is very recent. Their work is the first approach to show that it is possible to use only baseline HRCT scans to predict progression of idiopathic pulmonary fibrosis using artificial intelligence. In their paper they try to develop a novel predictive model for the radiological progression pattern of idiopathic pulmonary fibrosis using only baseline HRCT scans. First, they implemented a study design and having an expert radiologist contour region of interests (ROI) at baseline scans, depending on its progression status in follow-up visits. Then they integrated the feature selection with prediction by developing an algorithm using a wrapper method that combines quantum particle swarm optimization to select a small number of features with random forest to classify early patterns of progression. They compare their result with other popular wrappers and non-wrapper methods, i.e. smoothly clipped absolute deviation (SCAD), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), neural network (NNET). Their proposed model yields an overall accuracy rate of 82.1% which is superior to other feature selections and classification methods mentioned above.

Next paper titled "Idiopathic Pulmonary Fibrosis: Gender-Age-Physiology Index Stage for Predicting Future Lung Function Decline" [2] by Margaret L. Salisbury. In the paper they showed that patients with idiopathic pulmonary fibrosis (N = 657) were identified retrospectively at three tertiary referral centers, and baseline GAP stages were assessed. Mixed models were used to describe average trajectories of FVC and diffusing capacity of the lung for carbon monoxide (DLCO). Multivariable Cox proportional hazards models were used to assess whether declines in pulmonary function $\geq$ 10% in 6 months predict mortality after accounting for GAP stage. They found that over a 2-year period, GAP stage was not associated with differences in yearly lung function decline. After accounting for stage, a 10% decrease in FVC or DLCO over 6 months independently predicted death or transplantation. Patients with GAP stage 2 with declining pulmonary function experienced a survival profile similar to patients with GAP stage 3, with 1-year event-free survival of 59.3%. They came to a conclusion that baseline GAP stage predicted death or lung transplantation but not the rate of future pulmonary function decline.

A study has been done in [3] by Ana Adriana Trusculescu et al. titled Deep learning in interstitial lung disease. In their work they describe that interstitial lung diseases are a diverse group of disorders that involve inflammation and fibrosis of interstitium, with clinical, radiological, and pathological overlapping features. These are an important cause of morbidity and mortality among lung diseases. This review describes computer-aided diagnosis systems centered on deep learning approaches that improve the diagnostic of interstitial lung diseases. They highlighted the challenges and the implementation of important daily practice, especially in the early diagnosis of idiopathic pulmonary fibrosis (IPF). They developed a convolutional neuronal network (CNN) that could be deployed on any computer station and be accessible to non-academic centers is the next frontier that needs to be crossed.

In [4] Shudong Wang et al. classify lung cancer from CT images by deep residual neural networks with transfer learning strategy. They discuss about the accurate judgment of the pathological type of lung cancer is vital for treatment. Traditionally, the pathological type of lung cancer requires a histopathological examination to determine, which is invasive and time consuming. In their work, a novel residual neural network is proposed to identify the pathological type of lung cancer via CT images. Due to the low amount of CT images in practice, they explored a medical-to-medical transfer learning strategy. Specifically, a residual neural network is pre trained on public medical images dataset luna16, and then fine-tuned on their intellectual property lung cancer dataset collected in Shan-dong Provincial Hospital. Data experiments shows that their method achieves 85.71% accuracy in identifying pathological types of lung cancer from CT images and outperforming other models trained with 2054 labels. They show that their method performs better than AlexNet, VGG16 and DenseNet, which provides an efficient, non-invasive detection tool for pathological diagnosis.

In [5] Hyunkwang Lee et al. proposed an explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. They saw the improvements in image recognition via deep learning, machine-learning algorithms could eventually be applied to automated medical diagnoses that can guide clinical decision-making. However, these algorithms remain a ‘black box’ in terms of how they generate the predictions from the input data. Also, high-performance deep learning requires large, high-quality training datasets. They report the development of an understandable deep-learning system that detects acute intracranial haemorrhage (ICH) and classifies five ICH subtypes from unenhanced head computed-tomography scans. By using a dataset of only 904 cases for algorithm training, their system achieved a performance similar to that of expert radiologists in two independent test datasets containing 200 cases (sensitivity of 98% and specificity of 95%) and 196 cases (sensitivity of 92% and specificity of 95%). The system includes an attention map and a prediction basis retrieved from training data to enhance explainability, and an iterative process that mimics the workflow of radiologists.

In another work [6] Yutong Xie presented an knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT scan. The paper shows that the accurate identification of malignant lung nodules on chest CT is critical for the early detection of lung cancer, which also offers patients the best chance of cure. Deep learning methods have recently been successfully introduced to computer vision problems, although substantial challenges remain in the detection of malignant nodules due to the lack of large training datasets. In the paper, they propose a multi-view knowledge-based collaborative (MV-KBC) deep model to separate malignant from benign nodules using limited chest CT data. The model learns 3D lung nodule characteristics by decomposing a 3D nodule into nine fixed views. For each view, they construct a knowledge-based collaborative (KBC) sub model, where three types of image patches are designed to fine-tune three pre-trained ResNet-50 networks that characterize the nodules overall appearance, voxel and shape heterogeneity, respectively. They jointly use the nine KBC sub models to classify lung nodules with an adaptive weighting scheme learned during the error back propagation, which enables the MV-KBC model to be trained in an end-to-end manner. The penalty loss function is used for better reduction of the false negative rate with a minimal effect on the overall performance of the MV-KBC model. They tested their method on the benchmark LIDC-IDRI dataset and compared it to five state-of-the-art classification approaches. The results show that the MV-KBC model achieved an accuracy of 91.60% for lung nodule classification with an AUC of 95.70%. These results are markedly superior to the state-of-the-art approaches.

Chapter 3: Methodology

This chapter gives an overview of the different parts of our work chronologically. In this section we will discuss about the theories, techniques, and step by step workflow of our work.

3.1 Workflow

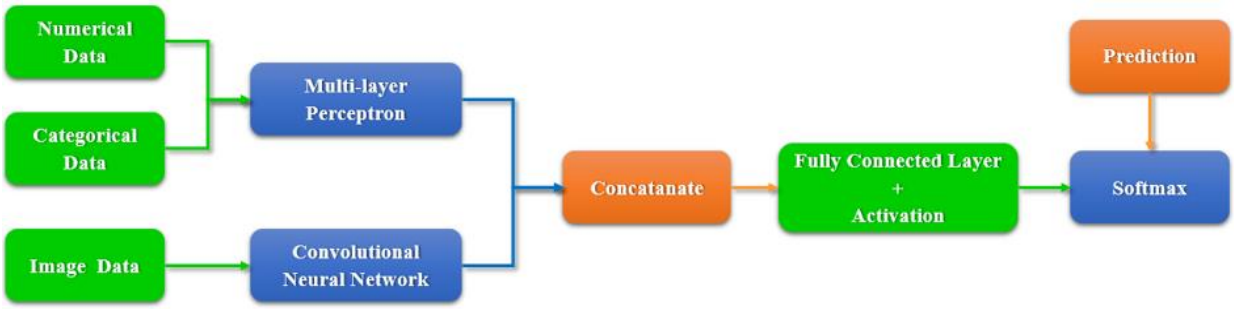

Figure 4: Overall workflow

This is a tentative workflow of our project. As we progress this workflow can change.

In this project our main objective is to predict the progression of idiopathic pulmonary fibrosis. But the prediction is notoriously difficult. So, we proposed an approach to build a model that can hopefully predict the progression of idiopathic pulmonary fibrosis.

Image Preprocessing is the most important step of working with image data. In our case the image data is the CT scan. During image preprocessing, we simultaneously prepare the images for our network and apply data augmentation to the training set.

We will use multilayer perceptron in order to process numerical and categorical data. Multilayer perceptrons train on a set of input-output pairs and learn to model the correlation between those inputs and outputs. In order to minimize error multilayer perceptron training involves adjusting the parameters, or the weights and biases of the model.

After that we will concatenate them and process them through fully connected layer. Then we will use softmax to predict the progression in the form of FVC value and we will find confidence value for every predicted FVC value. Softmax function, an activation function that turns numbers also known as logits into probabilities that sum to one. Softmax function outputs a vector that represents the probability distributions of a list of potential outcomes. Softmax turn logits into probabilities by taking the exponents of each output and then normalize each number by the sum of those exponents so the entire output vector adds up to one. It computes softmax cross entropy between logits and labels.

3.3 Software and Tools

The list of software and tools that we will be using throughout this project is given below

- **Python**
  - Overall Scripting.
- **PyTorch**
  - Multi-layer perceptron, ConvNet, Backpropagation, Optimization and others.
- **Scikit-learn**
- **Numpy, Pandas, Matplotlib**
  - Data Visualization.
- **Cuda**
  - Execution.

3.3 Dataset

We have collected our dataset from Open Source Imaging Consortium (OSIC) which is a non-profit, co-operative effort between academia, industry and philanthropy. The dataset contains a baseline chest CT scan and associated clinical information for a set of patients. A patient has an image acquired at time Week=0 and has numerous follow up visits over the course of approximately 1-2 years. The first problem we face with the data is that the relative timing of FVC measurements varies widely. The timing of the initial measurement relative to the CT scan and the duration to the forecasted time points differ for each patient. Features of the dataset are:

- **Patient**: a unique Id for each patient (also the name of the patient's DICOM folder).
- **Weeks**: the relative number of weeks pre/post the baseline CT (may be negative).
- **FVC**: the recorded lung capacity in ml.
- **Percent**: a computed field which approximates the patient's FVC as a percent of the typical FVC for a person of similar characteristics.
- **Age**: Patient's age.
- **Sex**: Patient's gender.
- **Smoking Status**

All the related information about metadata and data visualization are given below:

- Total 1549 patients.
- 176 patients are unique.
- 33,026 files/images, 176 folders/patients.
- 187.0 average files/images per patient.
- 1,018 max files/images per patient.

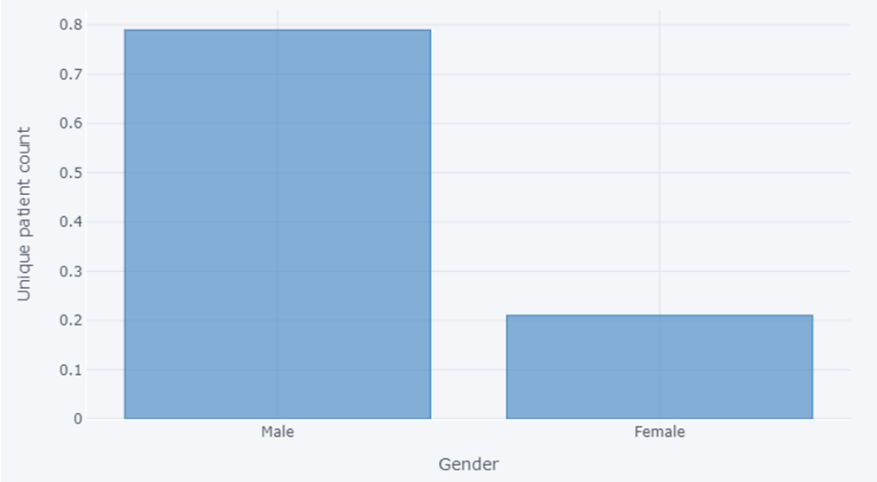

Figure 5: Gender distribution

Figure 5 shows that Almost 78% are Male and 22% are Female.

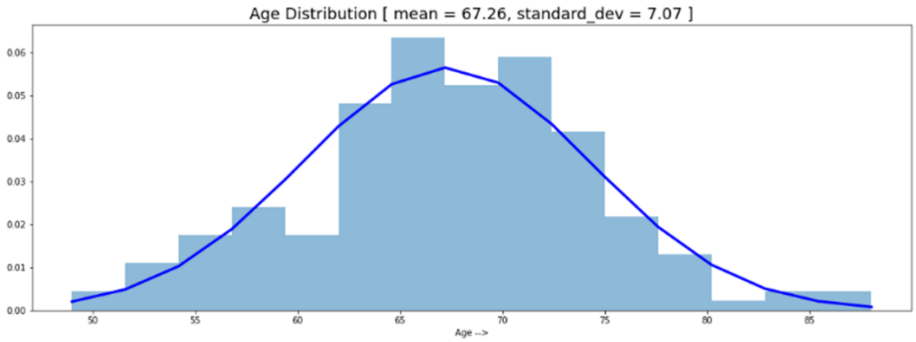

Figure 6: Age distribution

Figure 6 shows that age distribution varies from 49 years to 88 years.

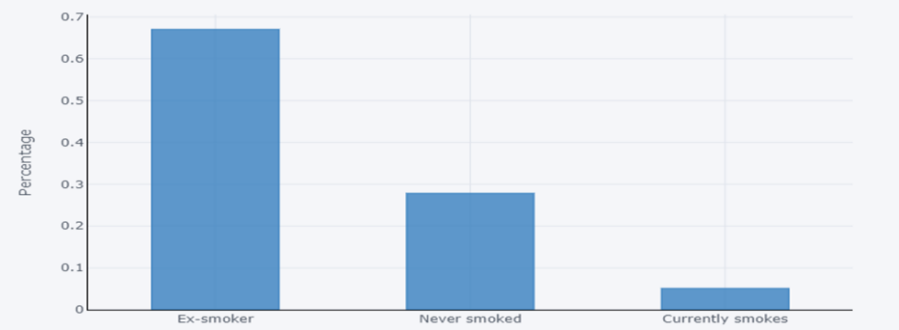

Figure 7: Smoking status distribution

Figure 7 shows that 67% of the patients are ex-smokers, 28% never smoked and 5% are still smoking

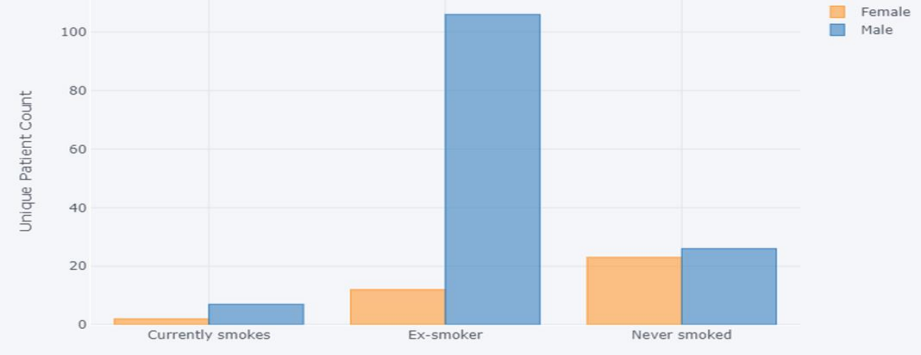

Figure 8: Gender vs smoking status

Figure 9: Age distribution of the patients based on Smoking Status

Figure 8 and 9 shows the gender vs smoking status and age distribution of the patients based on smoking status respectively.

3.4 Data Visualization and Preprocessing

The CT-scan captures information about the radiodensity of an object or tissue exposed to x-rays. A transversal slice of a scan is reconstructed after taking measurements from several different directions.

We need to transform to Hounsfield units as the spectral composition of the x-rays depends on the measurement settings like acquisition parameters and tube voltage. By normalizing to values of water and air (water has HU 0 and air -1000) the images of different measurements are becoming comparable. A CT-scanner yields roughly 4000 grey values that can't be captured by our eyes. This is why windowing is performed. This way the image is displayed in a HU range that suites most to the region of interest.

Transforming to Hounsfield Units:

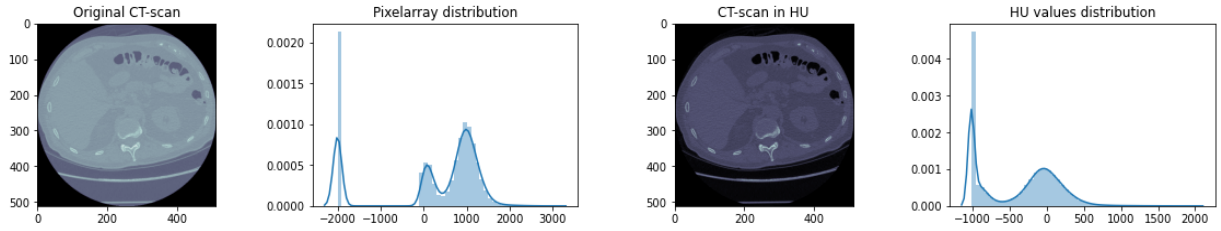

Figure 10: Original CT scan vs CT scan in hounsfield units

The voxel stands for the 3D-pixel that is given in a CT-scan. As far as we have discovered it is spanned by the 2d-plane of the pixelspacing attribute in x- and y-direction and the slice thickness in z-direction.

The pixelspacing attribute in the dicom files is an important one. It tells us how much physical distance is covered by one pixel. There are only 2 values that describe the x- and y-direction in the plane of a transversal slice. For one patient this pixelspacing is usually the same for all slices. But between patients the pixelspacing can differ due to personal or institutional preferences of doctors and the clinic and it also depends on the scanner type. Consequently, if we compare two images in the size of the lungs it does not automatically mean that the bigger one is really larger in the physical size of the organ.

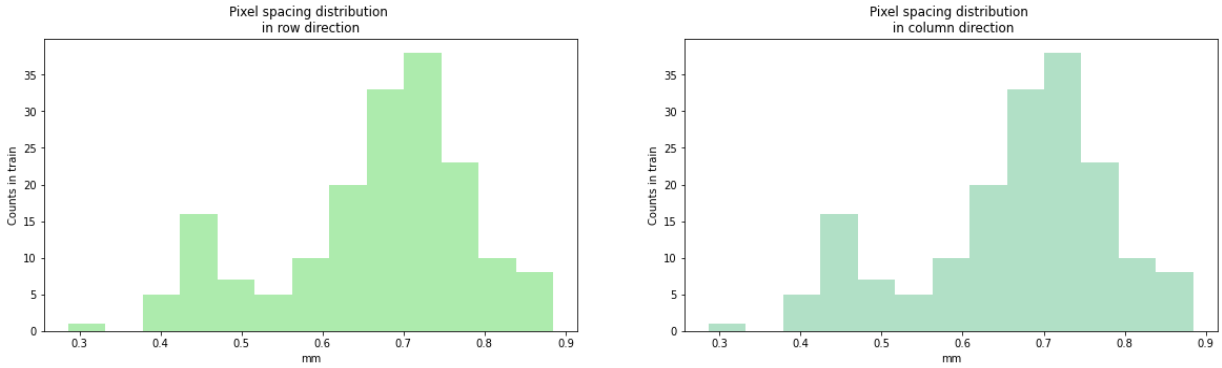

Figure 10: Pixelspacing distribution in row and column direction

The slice thickness tells us how much distance is covered in Z-direction by one slice. Figure 11 shows the distribution of it. Furthermore, the pixel array of raw values covers a specific area given by row and column values. Very thin slices allow more details to be shown. On the other hand, thick slices contain less noise but are more prone to artifacts.

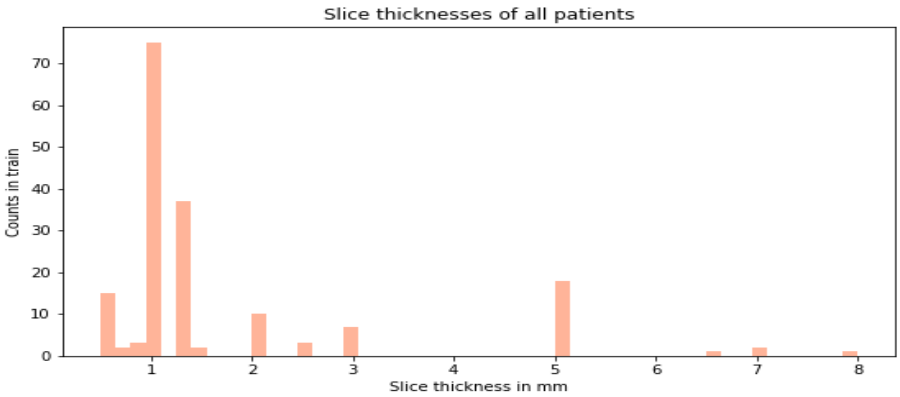

Figure 11: Slice thickness or pixelspacing distribution in Z-direction

Segmentation:

First of all, we are separating between potentially lung and air backgrounds. As we only like to segment the lungs, we need to remove the background. For this we will use morphological closing. Figure 12 shows how it looks like if we use morphological closing in our image with:

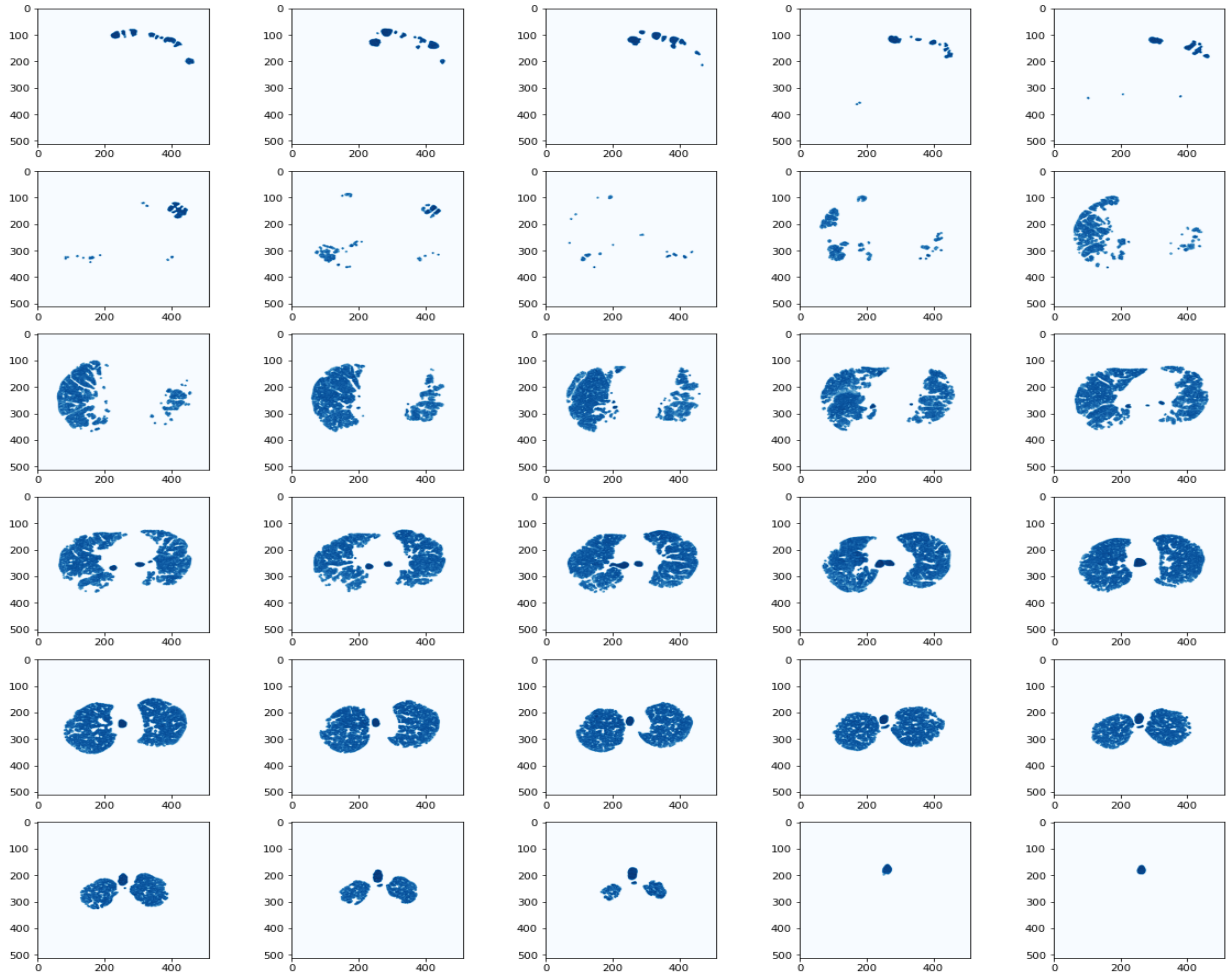

Figure 12: Segmented slice of lungs

To generate the data, we should take a look again at the different image sizes: For OSIC we have two major size groups and some minor outliers. Let's take a look at the sizes:

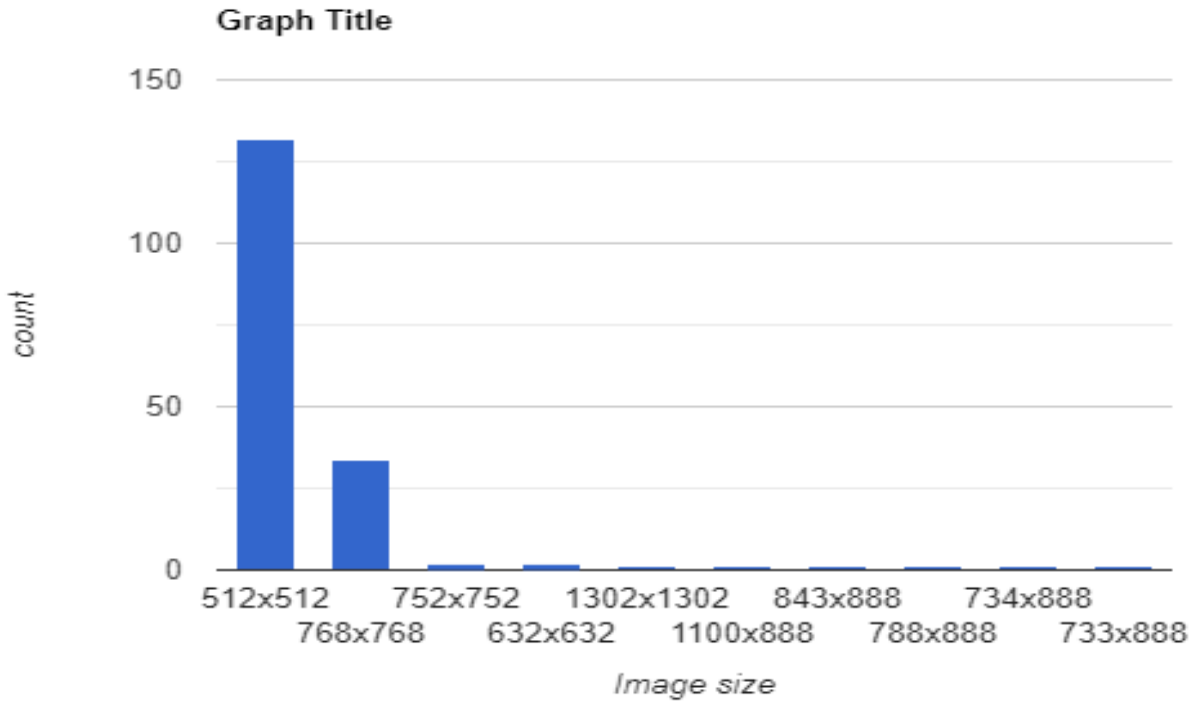

Figure 13: Image sizes vs patient count

As we can see we have 132 patients CT scan where the image size is 512×512. For image size of 768×768, total patient count is 34 and rest of are 1 or 2. As we have total 34000 image we need to resize all the image to a fixed size of 224×224, because we will use transfer learning and most of the transfer learning model uses the input size of 224×224.

3.5 Transfer Learning

As stated earlier we want to use transfer learning in our project. In this section we will discuss how to choose pre-trained model that is suitable for different problem and we also discuss about the architecture of different pre-trained model.

When people are repurposing a pre-trained model for their needs, they start by removing the original classifier, then add a new classifier that fits their purposes, and finally they have to fine-tune their model according to one of three strategies:

- **A. Train the entire model:** In this case, people use the architecture of the pre-trained model and train it according to their dataset. If someone learning the model from scratch, so they will need a large dataset and a lot of computational power.
- **B. Train some layers and leave the others frozen:** In pre-trained model, lower layers refer to general features which are problem independent, while higher layers refer to specific features which are problem dependent. Here, people can choose how much they want to adjust the weights of the network (a frozen layer does not change during training). Usually, if someone have a small dataset and a large number of parameters, they will leave more layers frozen to avoid overfitting. By contrast, if the dataset is large and the number of parameters is small, we can improve the model by training more layers.
- **C. Freeze the convolutional base:** This case corresponds to an extreme situation of the train/freeze trade-off. The main idea is to keep the convolutional base in its original form and then use its outputs to feed the classifier. In this case we are using the pre-trained model as a fixed feature extraction mechanism, which can be useful if someone is short on computational power, dataset is small, and/or pre-trained model solves a problem very similar to the one they want to solve.

Unlike Strategy C, whose application is straightforward, Strategy A and Strategy B require to be careful with the learning rate used in the convolutional part. The learning rate is a hyper-parameter that controls how much people adjust the weights of the network. When using a pre-trained model based on CNN, it's always a good practice to use a small learning rate because high learning rates increase the risk of losing previous knowledge. Assuming that the pre-trained model has been well trained, keeping a small learning rate will ensure that it don't distort the CNN weights too soon and too much.

There are perhaps a dozen or more top-performing pre-trained models for image recognition that can be downloaded and used as the basis for image recognition, classification and related tasks. Some of them are as follows:

- VGG (e.g. VGG16 or VGG19).
- GoogleNet (e.g. InceptionV3).
- Residual Network (e.g. ResNet34, ResNet50, ResNet101).

3.5.1 VGG16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the Oxford University in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". VGG16 refers to the fact that it has 16 layers that have some weights. 13 convolution and 3 fully connected layer. Total depth is 23 including input, convolution, pooling, fully connected layer and softmax. The dataset used for training, validation and testing is ImageNet dataset. It contains 1.2 million training images, 50,000 validation images, and 150,000 testing images and the size of the images is 256x256.

The input image size of VGG16 is 224x224. Therefore, the images have been down-sampled to a fixed resolution of 224x224. This is a pretty large network, and has a total of about 138.3 million parameters. The model achieves 92.7% test accuracy in ImageNet.

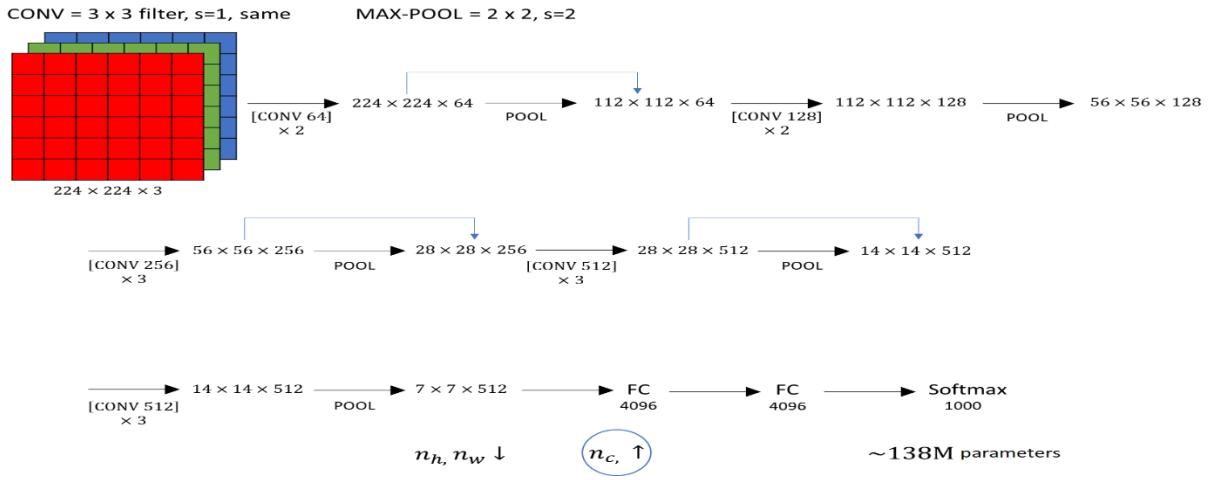

Figure 14: VGG16 architecture

Let's go through the architecture of VGG16.

- The first two layers are convolutional layers with 3×3 filters, and in the first two layers they use 64 filters that end up with a 224×224×64 volume because they are using same convolutions. So, this (CONV64)×2 represents that we have 2conv layers with 64 filters. The filters are always 3×3 with stride of 1 and they're always implemented with the same convolutions.
- Then, they use a pooling layer which will reduce height and width of a volume: it goes from 224×224×64 down to 112×112×64.
- Then, have a couple more conv layers. Here we use 128 filters and because we use the same convolutions, a new dimension will be 112×112×128.
- Then, a pooling layer is added so new dimension will be 56×56×128.
- 2conv layers with 256 filters
- The pooling layer
- A few more conv layers with 512 filters
- A pooling layer
- A few more conv layers with 512 filters
- A pooling layer
- At the end we have final 7×7×512 into Fullyconnected layer (FC) with 4096 units, and in a softmax output one of a 1000 classes. Which results in 138 illion parameters.

3.5.2 VGG19

VGG19 is similar to VGG1. It has 19 layers that have some weights consist of 16 convolution layer and 3 fully connected layer. The main difference of VGG19 with VGG16 is, it has three more layer of convolution than VGG16. So, Total depth is 26 including input, convolution, pooling,

fully connected layer and softmax. For three more layers of convolution Total number of parameters is around 143.6 million. VGG19 which is bigger than VGG16, but because VGG16 does almost as well as the VGG19 a lot of people use VGG16.

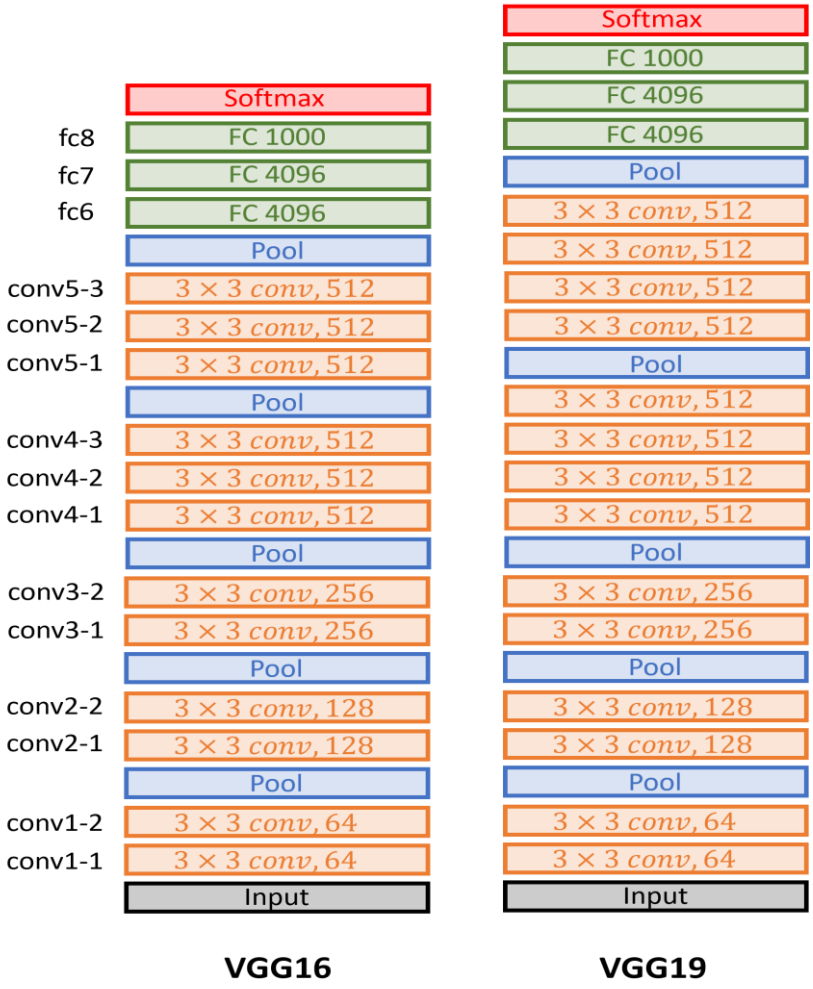

Figure 15: Comparison between VGG16 and VGG19 architecture

3.5.3 ResNet (34, 50, 100, 152)

Deep networks extract low, middle and high-level features and classifiers in an end-to-end multi-layer fashion, and the number of stacked layers can enrich the "levels" of features. When the deeper network starts to converge, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. The deterioration of training accuracy shows that not all systems are easy to optimize.

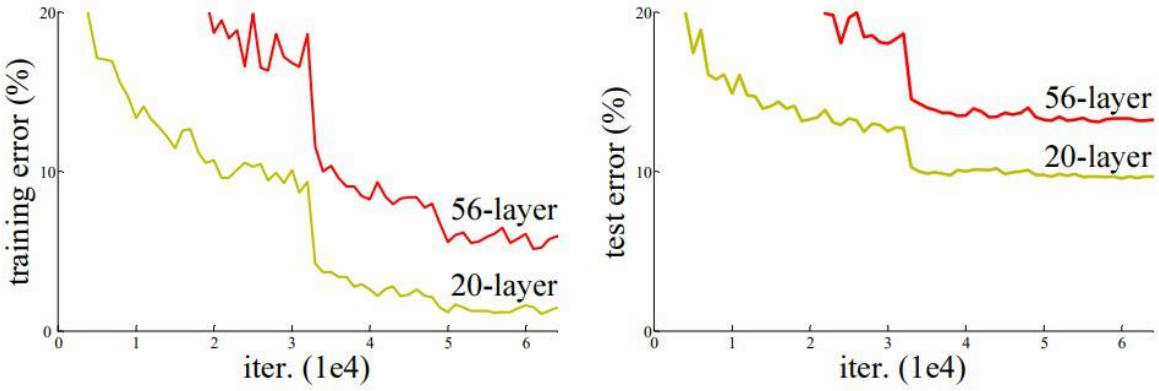

Figure 16: Training and testing error of deep network

To overcome this problem, Microsoft introduced a deep residual learning framework. Instead of hoping every few stacked layers directly fit a desired underlying mapping, they explicitly let these

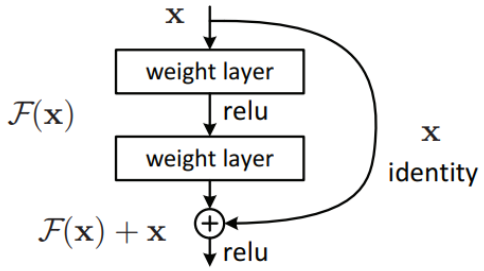

Figure 17: The residual block

layers fit a residual mapping. The formulation of F(x)+x can be realized by feedforward neural networks with shortcut connections. Shortcut connections are those skipping one or more layers. The shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers. Such residual part receives the input as an amplifier to its output. The dimensions usually are the same. Either way – no additional training parameters are used.

ResNet trained on ImageNet dataset. Its input size is 224x224. Total number of parameters of ResNet50 is around 25.6 million, for ResNet101 is 44.7 million and for ResNet152 is 60.4 million. ResNet50, Resnet101 and ResNet152 Achieved an accuracy of 92.1%, 92.8% and 93.1% respectively on the test image.

Deeper network such as ResNet50, ResNet101 use bottleneck layer to improve efficiency. Keeps the time complexity same as the two layered convolutions. Which Allows us to increase the number of layers as well as to converge much faster. 152-layer ResNet has 11.3 billion FLOPS while VGG-16/19 nets have 15.3/19.6 billion FLOPS

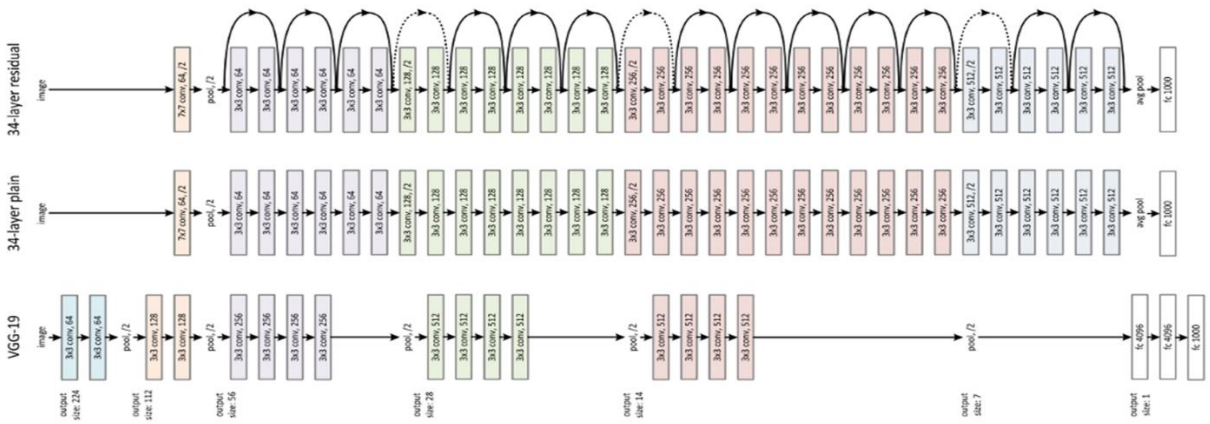

Figure 17: ResNet34 architecture in comparison with VGG19

The plain baselines, which is the middle figure are mainly inspired by the philosophy of VGG networks. The convolutional layers mostly have 3×3 filters and follow two simple rules. One is, for the same output feature map, the layers have the same number of filters and number two is, if the size of the features map is halved, the number of filters is doubled to preserve the time complexity of each layer.

It is worth noticing that the ResNet model has fewer filters and lower complexity than VGG networks. Based on the above plain network, a shortcut connection is inserted (Upper side figure) which turn the network into its counterpart residual version. The identity shortcuts can be directly used when the input and output are of the same dimensions as you can see from the solid line shortcut. When the dimensions increase (dotted line shortcuts) it considers two options. First is the shortcut performing identity mapping, with extra zero entries padded for increasing dimensions. This option introduces no additional parameter. Second is the projection shortcut is used to match dimensions (done by 1×1convolutions). For either of the options, if the shortcuts go across feature maps of two size, it performed with a stride of 2. Each ResNet block is either two layers deep (used in small networks like ResNet 34) or 3 layers deep (such as ResNet 50, 101, 152).

References

[1] Shi, Y., Wong, W. K., Goldin, J. G., Brown, M. S., & Kim, G. H. J. (2019). Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach. *Artificial intelligence in medicine*, *100*, 101709.

[2] Salisbury, M. L., Xia, M., Zhou, Y., Murray, S., Tayob, N., Brown, K. K., ... & Flaherty, K. R. (2016). Idiopathic pulmonary fibrosis: gender-age-physiology index stage for predicting future lung function decline. Chest, 149(2), 491-498.

[3] Trusculescu, A. A., Manolescu, D., Tudorache, E., & Oancea, C. (2020). Deep learning in interstitial lung disease—how long until daily practice. European Radiology, 1-8.

[4] Wang, S., Dong, L., Wang, X., & Wang, X. (2020). Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy. Open Medicine, 15(1), 190-197.

[5] Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S. H., Guerrier, C. E., ... & Gonzalez, R. G. (2019). An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nature Biomedical Engineering, 3(3), 173.

[6] Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., & Cai, W. (2018). Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE transactions on medical imaging, 38(4), 991-1004.

[7] Walsh, S. L., Calandriello, L., Silva, M., & Sverzellati, N. (2018). Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. The Lancet Respiratory Medicine, 6(11), 837-845.