

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 # 1621529042

Faculty Advisor

Dr. Tanzilur Rahman

Assistant Professor

ECE Department

Fall 2020

LETTER OF TRANSMITTAL

December 2020

To

Dr. Mohammad Rezaul Bari

Associate Professor & Chair

Department of Electrical & Computer Engineering

North South University, Dhaka

Subject: **Submission of Capstone Project on “Prediction of Idiopathic Pulmonary Fibrosis Progression Using Deep Learning”.**

Dear Sir,

With due respect, we would like to submit our Capstone Project Report on “Prediction of Idiopathic Pulmonary Fibrosis Progression Using Deep Learning” as a part of our B.Sc. program. The report deals with Prediction of Idiopathic Pulmonary Fibrosis Progression. It helps patients and their families to understand their prognosis when they are first diagnosed with this incurable lung disease. We tried our level best to make the report meaningful and informative. This project was very much valuable to us as it helped us gain experience from practical field and apply in real life. It was a great learning experience for us. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you are kind enough to receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

.....
Shazzad Hasan
ECE Department
North South University, Bangladesh

.....
Md. Tanzim Hossain
ECE Department
North South University, Bangladesh

.....
Md. Saidur Rahman
ECE Department
North South University, Bangladesh

APPROVAL

The capstone project entitled “**Prediction of Idiopathic Pulmonary Fibrosis Progression Using Deep Learning**” by **Shazzad Hasan (ID # 1530604043)**, **Md. Tanzim Hossain (ID # 1620776042)** and **Md. Saidur Rahman (ID # 1621529042)**, is approved in partial fulfillment of the requirement of the Degree of Bachelor of Science in Computer Science and Engineering on December, 2020 and has been accepted as satisfactory.

Supervisor’s Signature

.....
Dr. Tanzilur Rahman
Assistant Professor
Department of Electrical & Computer Engineering
North South University
Dhaka, Bangladesh.

Department Chair’s Signature

.....
Dr. Mohammad Rezaul Bari
Associate Professor & Chair
Department of Electrical & Computer Engineering
North South University
Dhaka, Bangladesh.

DECLARATION

This is our truthful declaration that the “Capstone Project Report” we have prepared is not a copy of any “Capstone Project Report” previously made by any other team. We also express our honest confirmation in support of the fact that the said “Capstone Project Report” has neither been used before to fulfill any other course related purpose nor it will be submitted to any other team or authority in future. Any material reproduced in this project has been properly acknowledged.

Students’ name & Signature

.....
Shazzad Hasan
ECE Department
North South University, Bangladesh

.....
Md. Tanzim Hossain
ECE Department
North South University, Bangladesh

.....
Md. Saidur Rahman
ECE Department
North South University, Bangladesh

ACKNOWLEDGEMENT

First of all, we wish to express our gratitude to the Almighty for giving us the strength to perform our responsibilities and complete the report.

The capstone project program is very helpful to bridge the gap between the theoretical knowledge and real life experience as part of Bachelor of Science (B.Sc.) program. This report has been designed to have a practical experience through the theoretical understanding.

We also acknowledge our profound sense of gratitude to all the teachers who have been instrumental for providing us the technical knowledge and moral support to complete the project with full understanding.

We would like to convey our gratitude to our faculty **Dr. Tanzilur Rahman** for his stimulating inspiration, kind guidance, valuable suggestions, sagacious advice and kind cooperation throughout the period of work undertaken, which has been instrumented in the success of our project. At this level of understanding it is often difficult to understand the wide spectrum of knowledge without proper guidance and advice. His suggestions & guidance have made the report a good manner.

We thank our **friends and family** for their moral support to carve out this project and always offer their support.

ABSTRACT

Idiopathic Pulmonary fibrosis is a progressive lungs disease which usually gets worse over time. Once this disease damages the lungs, it cannot be cured totally. But early detection and proper diagnosis can help to keep this disease in control. It causes scarring in the lungs over time. As an effect, people face breathing difficulty. It can cause shortness of breath, even at rest. The general causes of pulmonary fibrosis can be exposure to toxic element like coal dust, asbestos fibres, silica dust, hard metal dusts etc. But in majority of the cases, the doctor cannot figure out the exact cause of this disease. That's why this disease is termed as Idiopathic Pulmonary Fibrosis. In near future, early diagnosis of pulmonary fibrosis should be possible. Deep learning model is helping to use the human resources efficiently and it is also reducing the expenses spent on the social and healthcare aspects of this deadly disease. In this project our job is to build a model that can predict the progression of idiopathic pulmonary fibrosis. We complete this project in two parts. in the first part we work with image data along with some portion of categorical data to build a model using transfer learning. In the second part we use the model and to generate new data frame to work with. Using that data frame with some modification and calculation we build the final model to predict the progression. We have trained 6 different model, which include 4 different type of Efficientnet namely B0, B1, B2, B4. And ResNet50 and VGG16. Along with these model we also use quantile regression to build our final model. We have evaluated our trained model on a modified version of the Laplace Log Likelihood and we have achieved the evaluation score as high as to -6.6767 which is better than any other existing model.

Contents

1	Introduction	12
1.1	Lung	12
1.2	Idiopathic Pulmonary Fibrosis	14
1.2.1	Symptoms of Idiopathic Pulmonary Fibrosis	15
1.2.2	Causes and Risks of Idiopathic Pulmonary Fibrosis	16
1.2.3	Diagnosis of Idiopathic Pulmonary Fibrosis	17
1.2.4	Treatments of Idiopathic Pulmonary Fibrosis	18
1.2.5	Complications of Idiopathic Pulmonary Fibrosis	19
1.3	CT scan	20
1.4	Motivation	21
1.5	Aim and Objective	22
2	Background Study	23
3	Dataset	27
3.1	Description	27
3.2	Visualization	28
3.3	Preprocessing	35

4	Methodology	40
4.1	Software and Tools	41
4.2	Transfer Learning	41
4.2.1	VGG16	43
4.2.2	VGG19	43
4.2.3	ResNet (34, 50, 100, 152)	44
4.2.4	EfficientNet	47
4.3	Quantile Regression	49
4.4	Evaluation Metric: Laplace Log Likelihood	50
5	Experiments & Results	52
5.1	Overview	52
5.2	Result	56
6	Discussion	62
6.1	Future Work	62
6.2	Challenges Limitations	62
6.3	Conclusion	63

List of Figures

1.1	Lung	12
1.2	Lung with idiopathic pulmonary fibrosis	14
1.3	CT scan	20
3.1	Gender Distribution	29
3.2	Distribution of the Smoking Status column in the Unique Patient Set	30
3.3	Gender vs Smoking Status	30
3.4	Age Distribution in Training Set	30
3.5	FVC Distribution in Training Set	31
3.6	Week Distribution in Training Set	31
3.7	Percent Distribution in Training Set	31
3.8	Week vs Age Distribution	32
3.9	FVC vs Age Distribution	32
3.10	FVC vs Week Distribution Over Smoking Status	32
3.11	Age vs Percent Distribution Over Smoking Status	33
3.12	FVC vs Percent Distribution	33
3.13	Distribution of Smoking Status Over Percentage	33

3.14	Age Distribution of Male & Female Patient	34
3.15	Age Distribution With Respect to Smoking Habit	34
3.16	Pixel Array Distribution	36
3.17	Original CT-scan vs CT-scan in Hounsfield Unit	36
3.19	Physical Area & Slice Volume Covered by a Single CT-Scan	37
3.18	Pixel Spacing Distribution in Row and Column Direction .	37
3.20	Smallest & largest CT-slice volume	38
3.21	Image Sizes vs Patient Count	38
3.22	Segmented Slices of Lungs	39
4.1	VGG16 architecture	43
4.2	Comparison Between VGG16 and VGG19 Architecture . .	44
4.3	Training and testing error of deep network	45
4.4	The residual block	45
4.5	ResNet34 architecture in comparison with VGG19	46
4.6	Comparison of Different Scaling Methods.	47
4.7	Performance of EfficientNet	48
5.1	Model architecture using image data and partial categorical data	53
5.2	Final model architecture	55
5.3	Training & Validation Loss	59
5.4	Training loss curve after smoothing	60
5.5	Validation loss curve after smoothing	61

List of Tables

5.1	Example submission table	54
5.2	Example of starting week FVC for each patient	54
5.3	Updated data table	55
5.4	Comparison of Training Loss And Validation Loss with Different Model	57
5.5	Comparison of score of different models	58
5.6	Example submission table	58

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

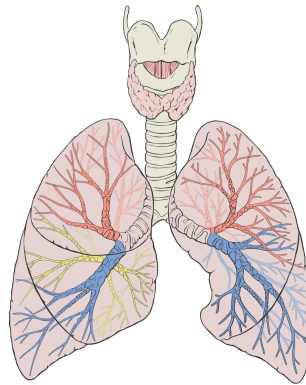


Figure (1.1) Lung

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.

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.

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.

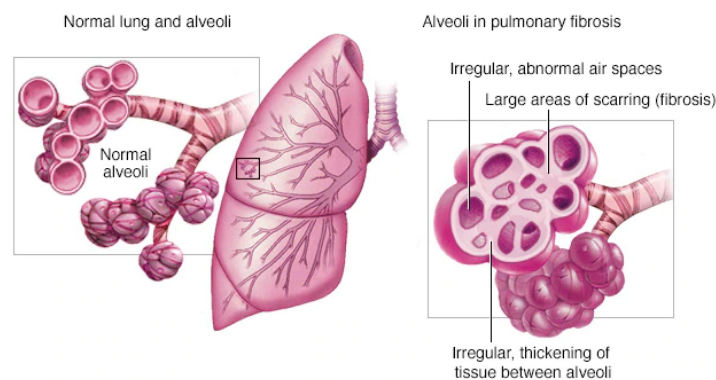


Figure (1.2) Lung with idiopathic pulmonary fibrosis

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 dyspnea, 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 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 normal 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



Figure (1.3) CT scan

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.

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

Background Study

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” [7] 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 abso-

lute 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” [6] 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 [10] 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 [11] 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 [3] 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 [12] 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.

A paper titled Prediction Analysis of Idiopathic Pulmonary Fibrosis Progression from OSIC Dataset [4] by work with the same dataset as ours. Their objective is to analyse and compare the performance of various machine learning models by predicting the final forced volume capacity measurements for each patient and a confidence value. Their main goal is to deploy the model on any computer to predict a patient's severe condition regarding lungs function which is based on a CT scan of the lungs of the patients. Lung function is checked out based on a spirometer output that measures the forced vital capacity (FVC) of the lungs. To conduct the research the mainly followed three strategy which are Multiple Quantile Regression, Ridge Regression and ElasticNet Regression. They also discussed the result that they have found by using those three strategy.

Chapter 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.

3.1 Description

Our dataset contains two kinds of data.

1. **HRCT scans**
2. **Tabular data**

Both kinds of data are associated with 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 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 tabular data 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:** Patient's smoking status.

All the related information about tabular data 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.

3.2 Visualization

In this section we will visualize our dataset from different point of view. The total number of entry in the dataset are 1549, among those 176 are unique. Figure 3.1 shows the male female distribution of the dataset. There are 139 male patient and 37 female patient in the dataset. The ratio of the male female data are 78% and 22%. Figure 3.2 shows that the dataset contains 118 ex-smokers data which is 67%, 18 patients data who never smoked which is 28% and 9 patients data who are currently smoking which is 5% of the total dataset. The above data shows that among the patients ex-smokers are higher in number. Figure 3.3 represents the combination of both figure 3.1 and figure 3.2. Among the patient who are ex-smoker 106 of them are male and 12 of them are female. 26 male patient and 23 female patient never smoked while 7 male patient and 2 female patient currently smokes. As of figure 3.4 the number of patients is higher within the age of 62-73. The force vital capacity (FVC), is the

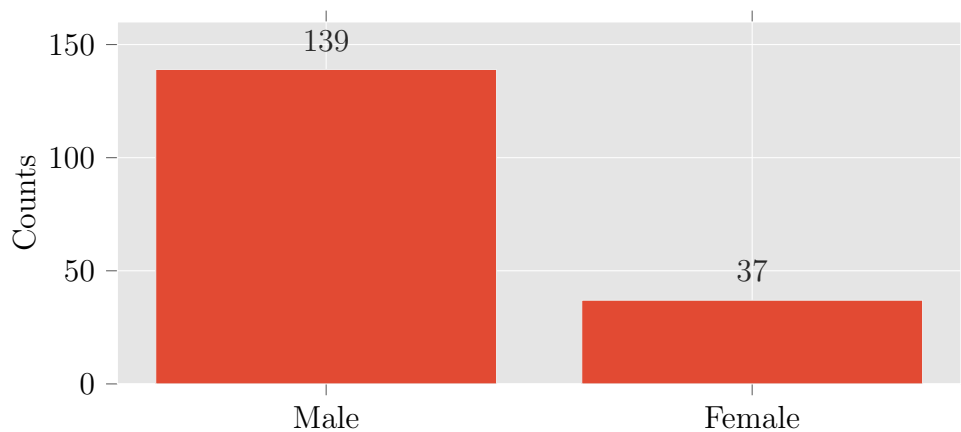


Figure (3.1) Gender Distribution

volume of air exhaled. The value of FVC recorded in ml. Figure 3.5 represents histogram of the fvc value in training set. The histogram of figure 3.6 shows the number of patients in weeks pre/post baseline CT-scan in training set. Negative weeks for pre baseline CT-scan. We can see that the density is very higher close to the baseline Ct-scan. Figure 3.7 shows the percent distribution of the training set. Figure 3.8 demonstrate the age versus week distribution for male and female patient. There are some negative values for Weeks. Because Weeks is the relative number, here negative values represents weeks of pre baseline CT-scan. Positive values indicates post baseline CT-scan. Blue scatter dots shows male patients and red dots are females. The distribution of figure 3.9 shows the relation between and age. From this data we can say that male have higher FVC value with respect to female. The scatter plots of figure 3.10 FVC value over number of weeks with respect to smoking status. Person never smoked has FVC lower than smoker. Some Ex-smoker have very high FVC. figure 3.11 shows the relationship between age over percent with respect to smoking status. Percent is very much related to Fvc value. As we can see from the figure 3.12 that higher FVC value relates to higher percent value. Figure 3.13 shows the co-relation between smoking habit and percent. For ex-smoker percent value is high at the range of 50-90, for the patients who never smoked percent value is high at the range of 50-70 and for current smokers percent value is high at 70-100. Figure 3.14 and 3.15 represents the the age distribution with respect to gender and smoking habit.

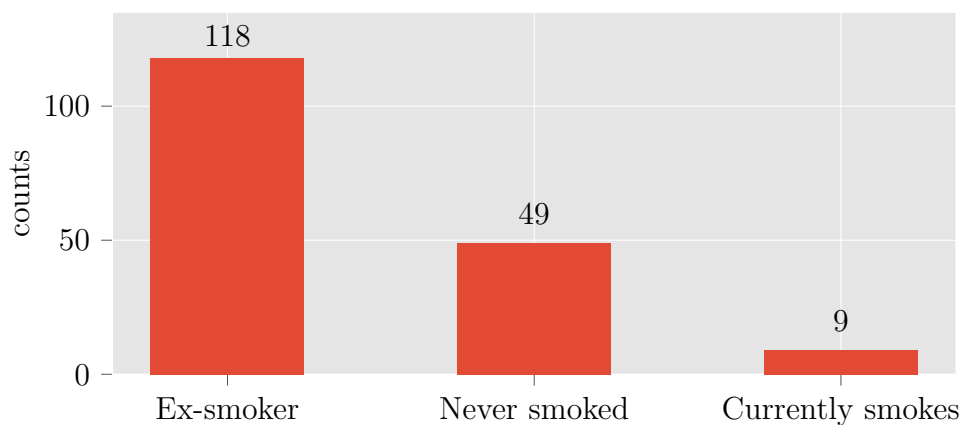


Figure (3.2) Distribution of the Smoking Status column in the Unique Patient Set

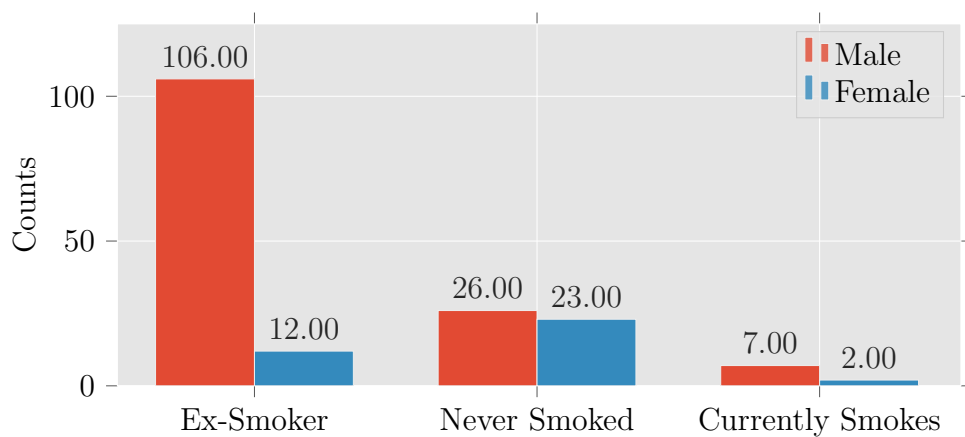


Figure (3.3) Gender vs Smoking Status

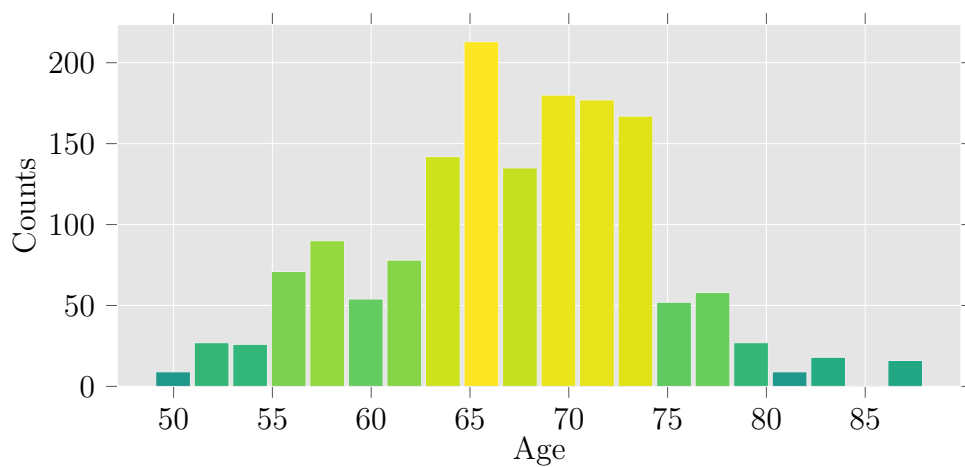


Figure (3.4) Age Distribution in Training Set

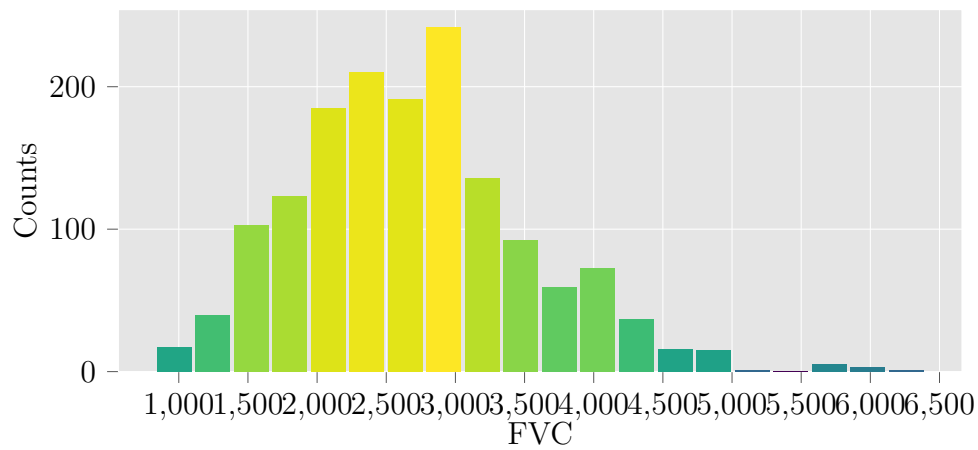


Figure (3.5) FVC Distribution in Training Set

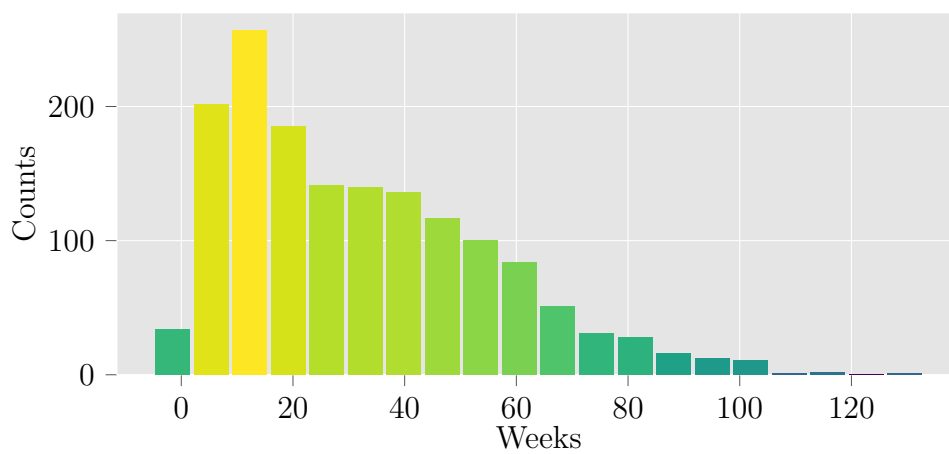


Figure (3.6) Week Distribution in Training Set

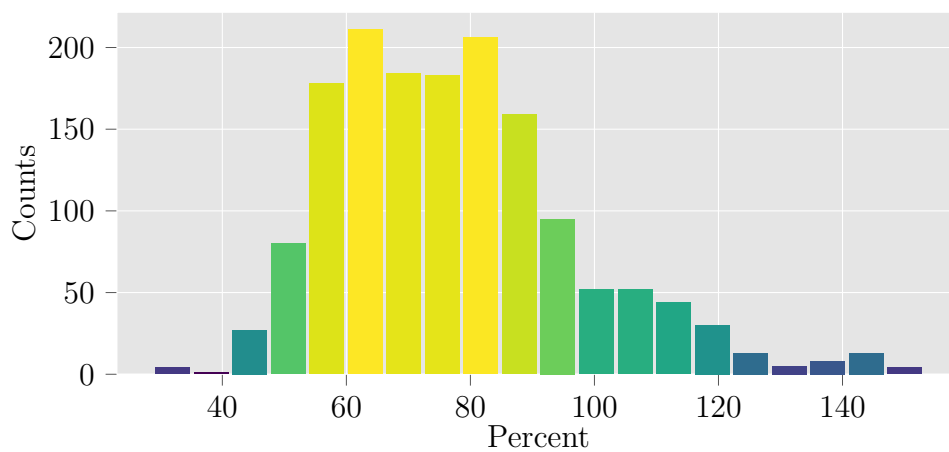


Figure (3.7) Percent Distribution in Training Set

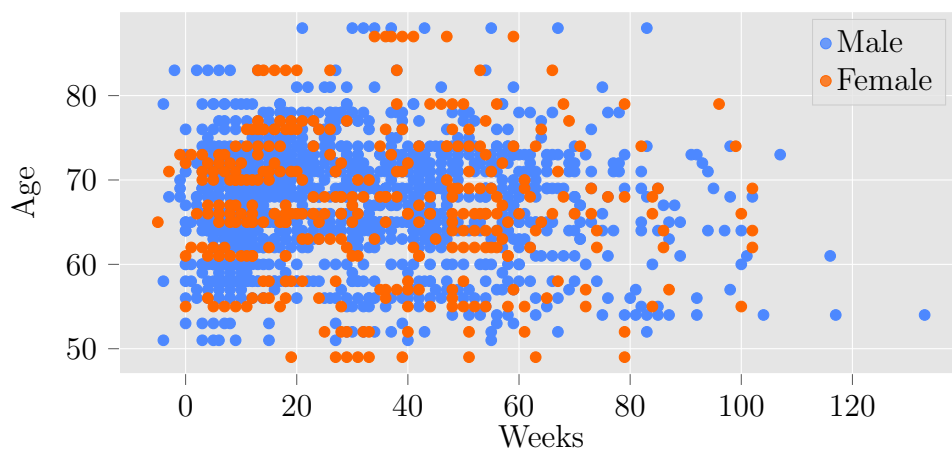


Figure (3.8) Week vs Age Distribution

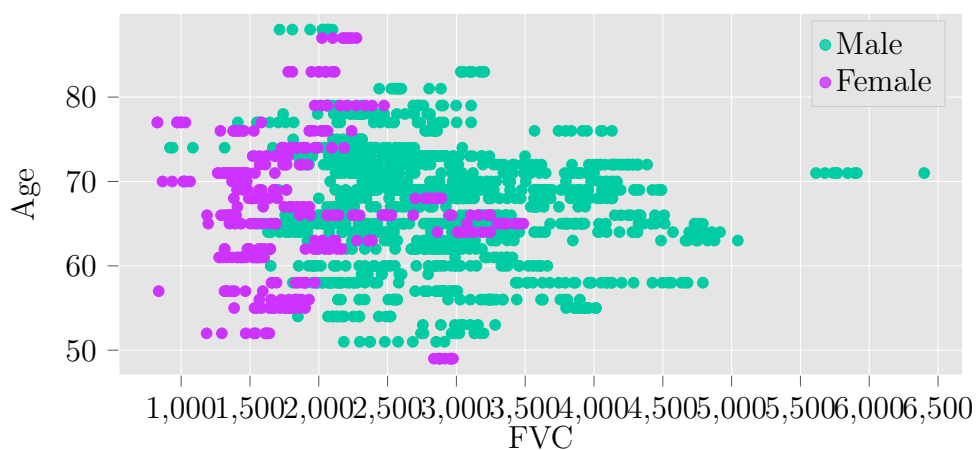


Figure (3.9) FVC vs Age Distribution

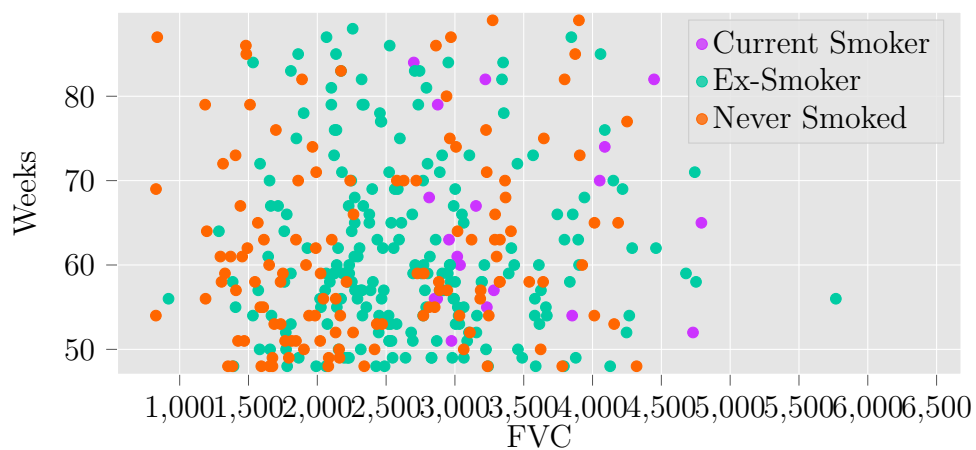


Figure (3.10) FVC vs Week Distribution Over Smoking Status

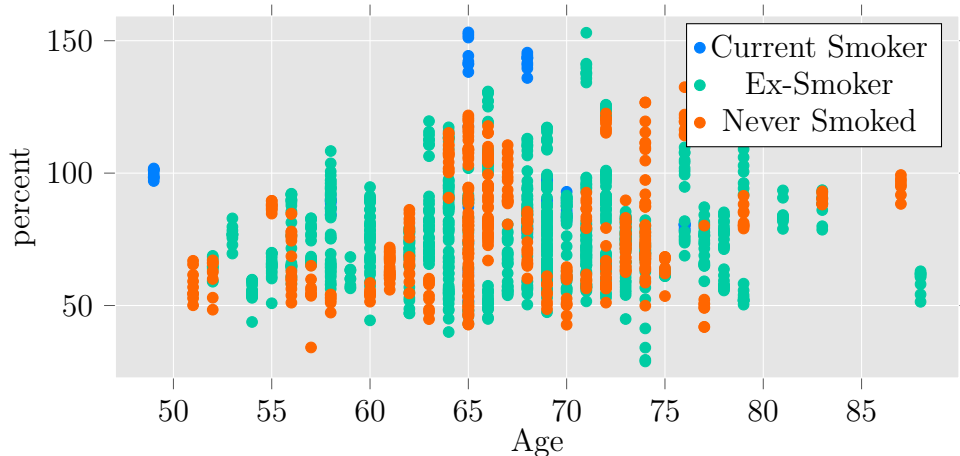


Figure (3.11) Age vs Percent Distribution Over Smoking Status

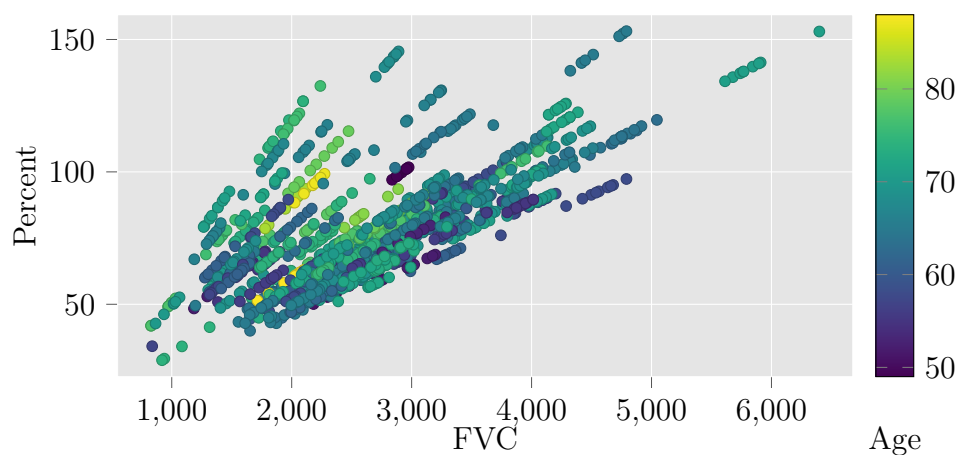


Figure (3.12) FVC vs Percent Distribution

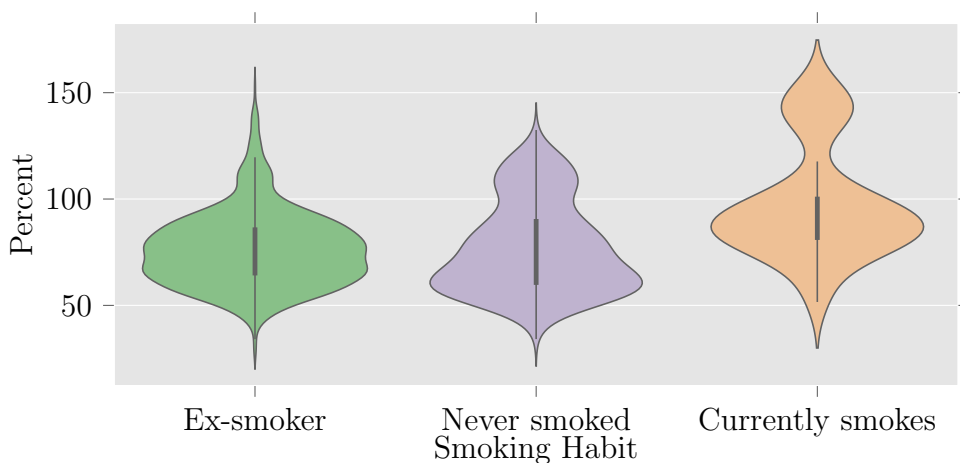


Figure (3.13) Distribution of Smoking Status Over Percentage

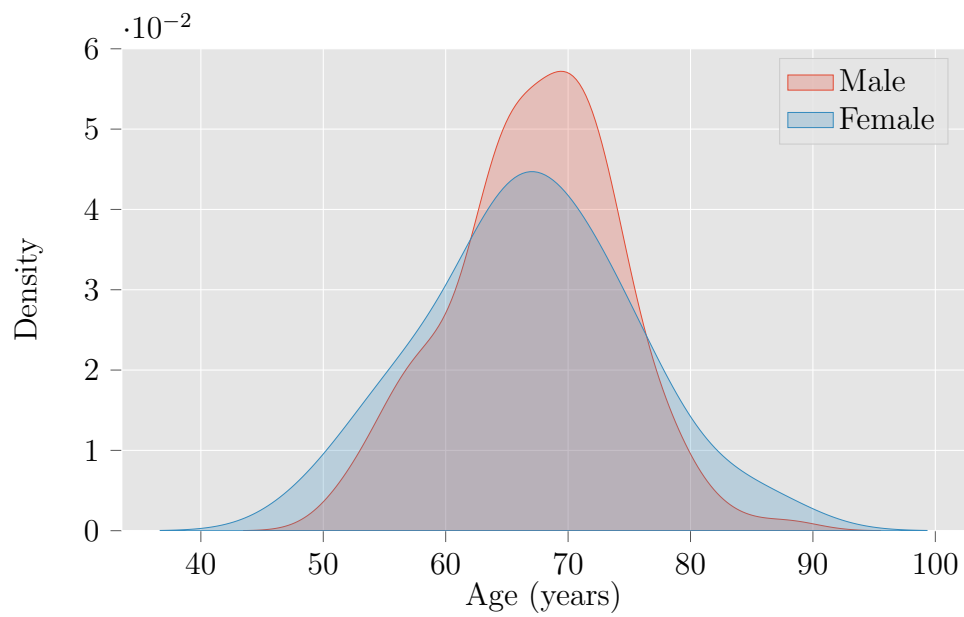


Figure (3.14) Age Distribution of Male & Female Patient

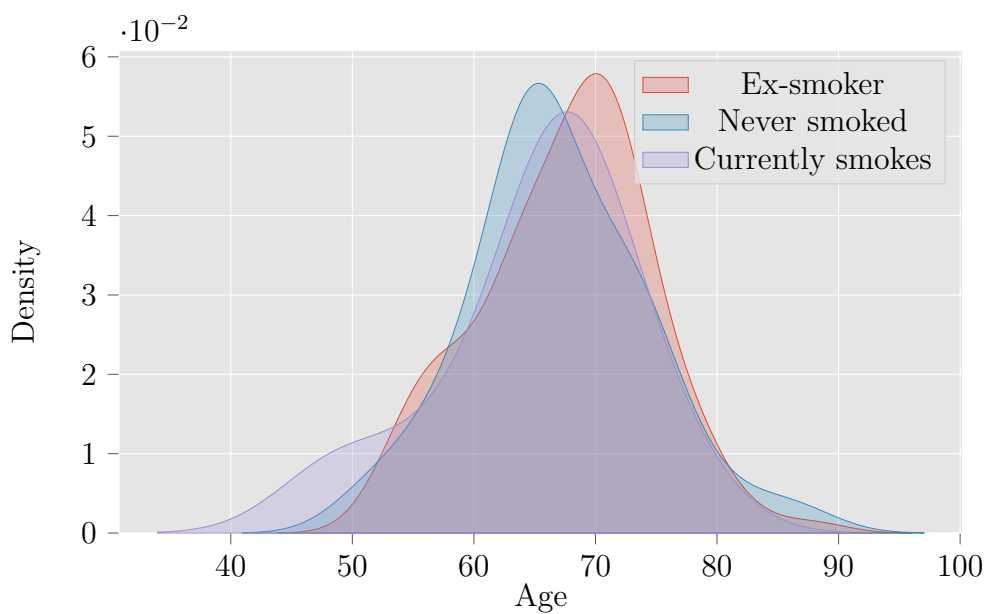


Figure (3.15) Age Distribution With Respect to Smoking Habit

3.3 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 the CT-scan file 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. Figure 3.17 shows the transformation from original CT-scan image to CT- scan in hounsfield unit.

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. Figure 3.18 shows the pixel spacing distribution in row and column direction. 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.

The slice thickness tells us how much distance is covered in Z-direction by one slice. Figure 3.19 shows the distribution of physical area & slice volume covered by a single ct-scan. 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.

From figure 3.20 we can see that the large one has a lot of useless

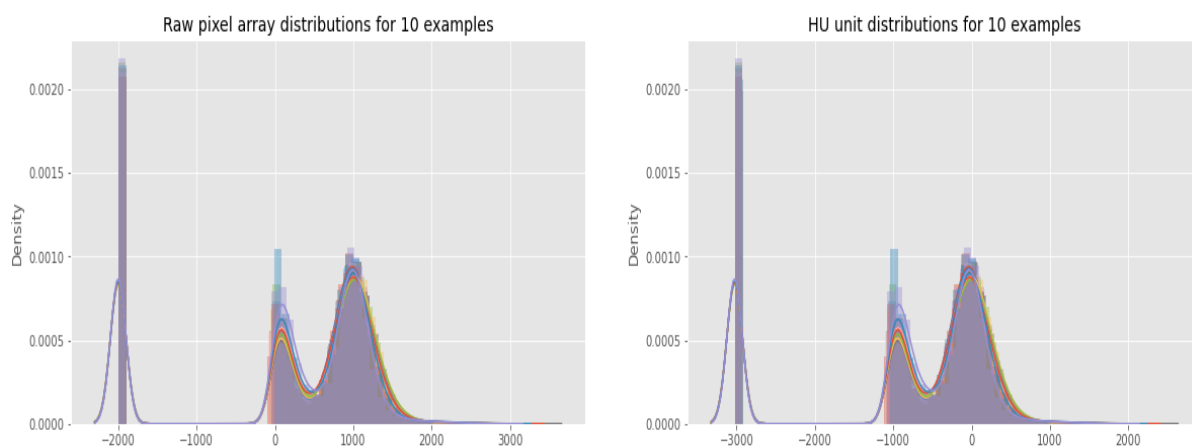


Figure (3.16) Pixel Array Distribution

region covered. In the second image with the large area the outside region of the scanner tube is not set to the value of air but rather to some value in the middle of the range -1000 to 1000.

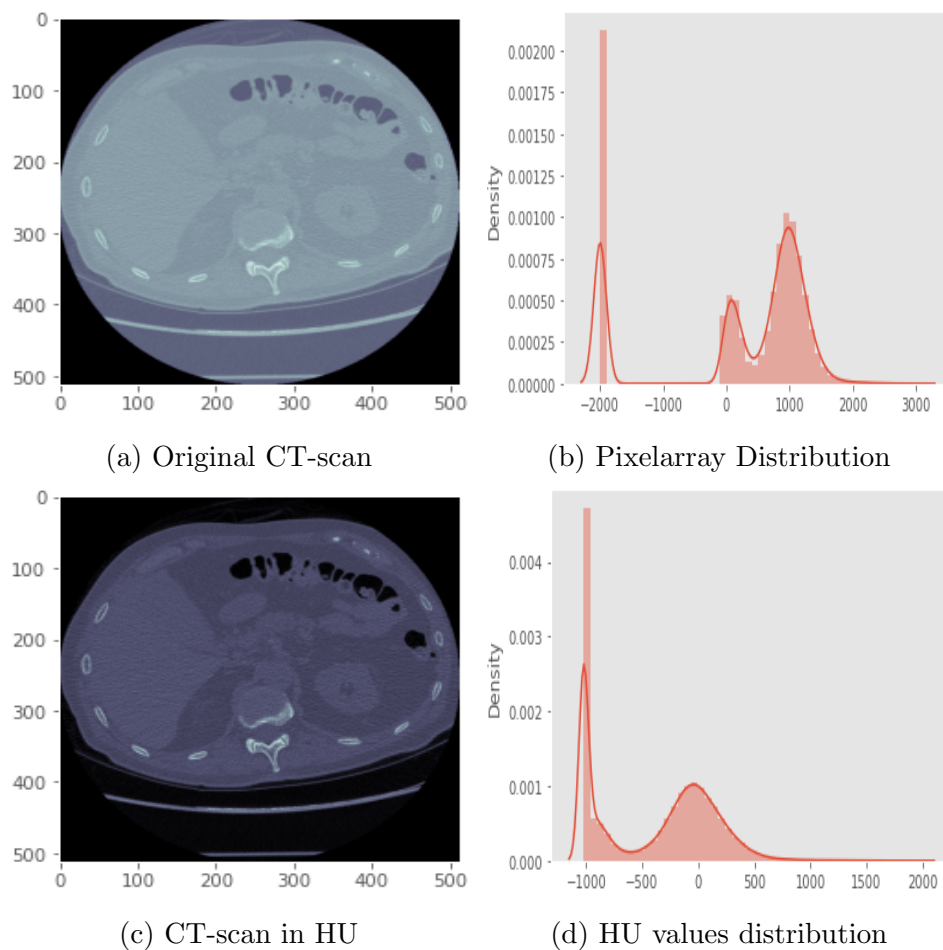


Figure (3.17) Original CT-scan vs CT-scan in Hounsfield Unit

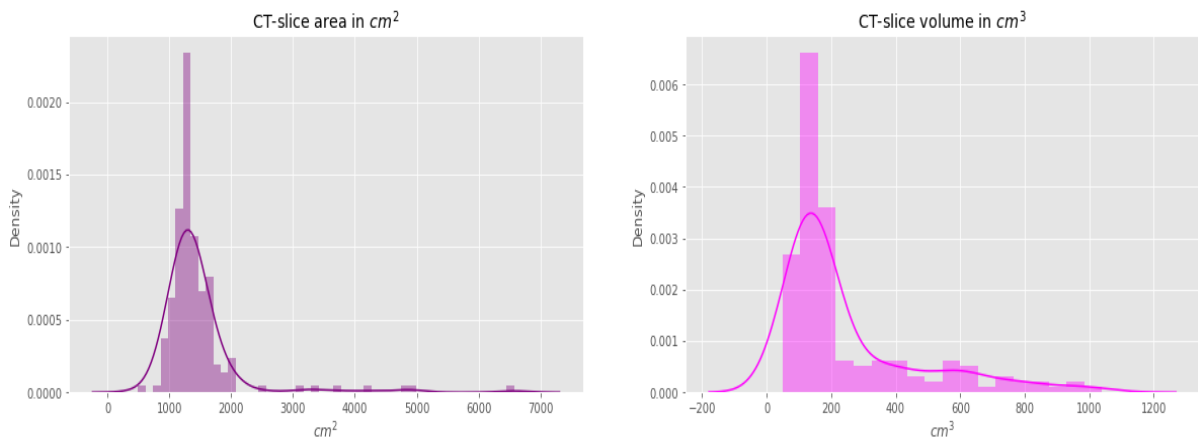


Figure (3.19) Physical Area & Slice Volume Covered by a Single CT-Scan

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 3.22 shows how it looks like if we use morphological closing in our image.

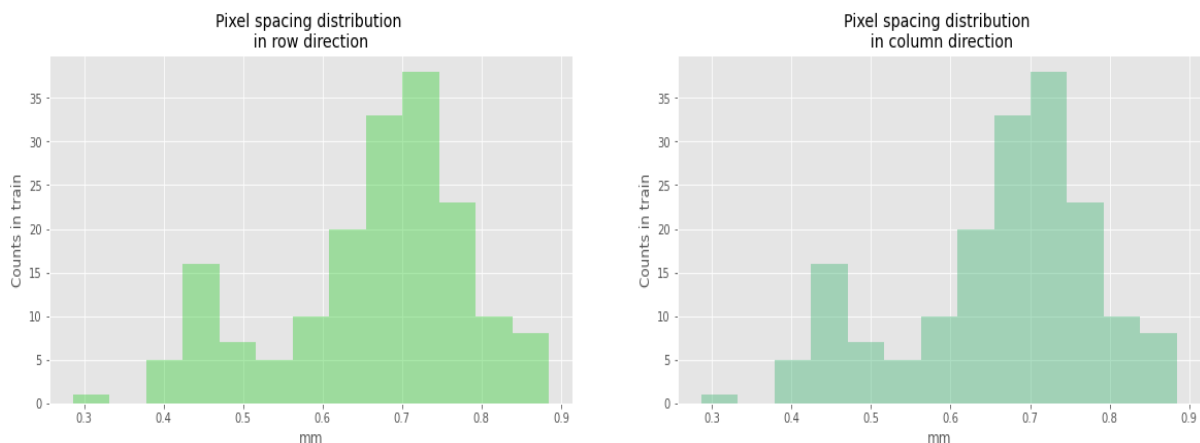


Figure (3.18) Pixel Spacing Distribution in Row and Column Direction

In order to generate pre-processed data, we should take a look again at the different image sizes. For our data we have two major size groups and some minor outliers. From figure 3.21 we can see that 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 512×512 , because the largest chunk of data are of that size. So, it will take less time to generate the data.

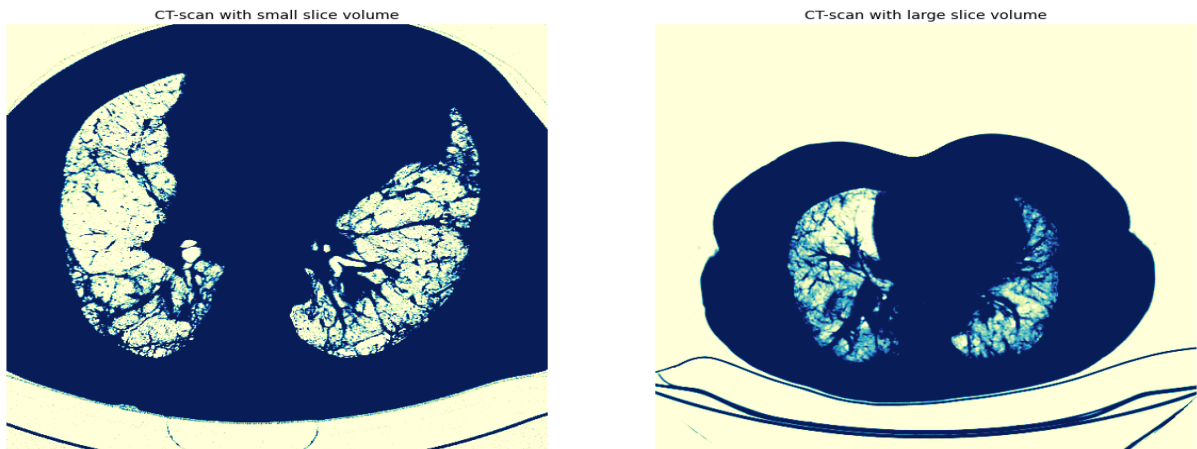


Figure (3.20) Smallest & largest CT-slice volume

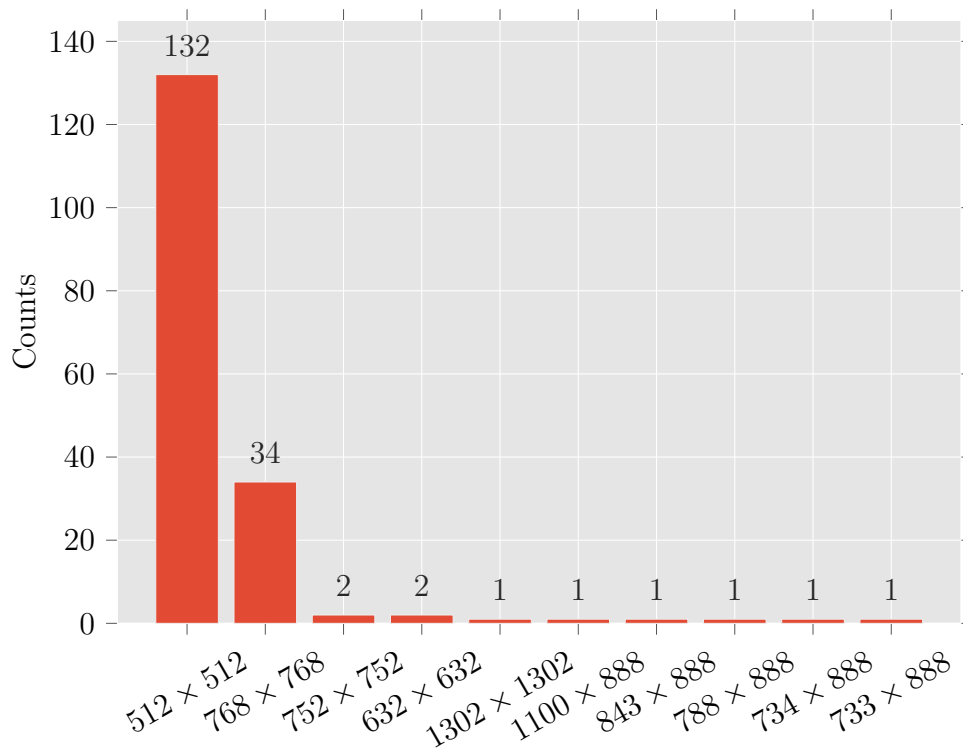


Figure (3.21) Image Sizes vs Patient Count

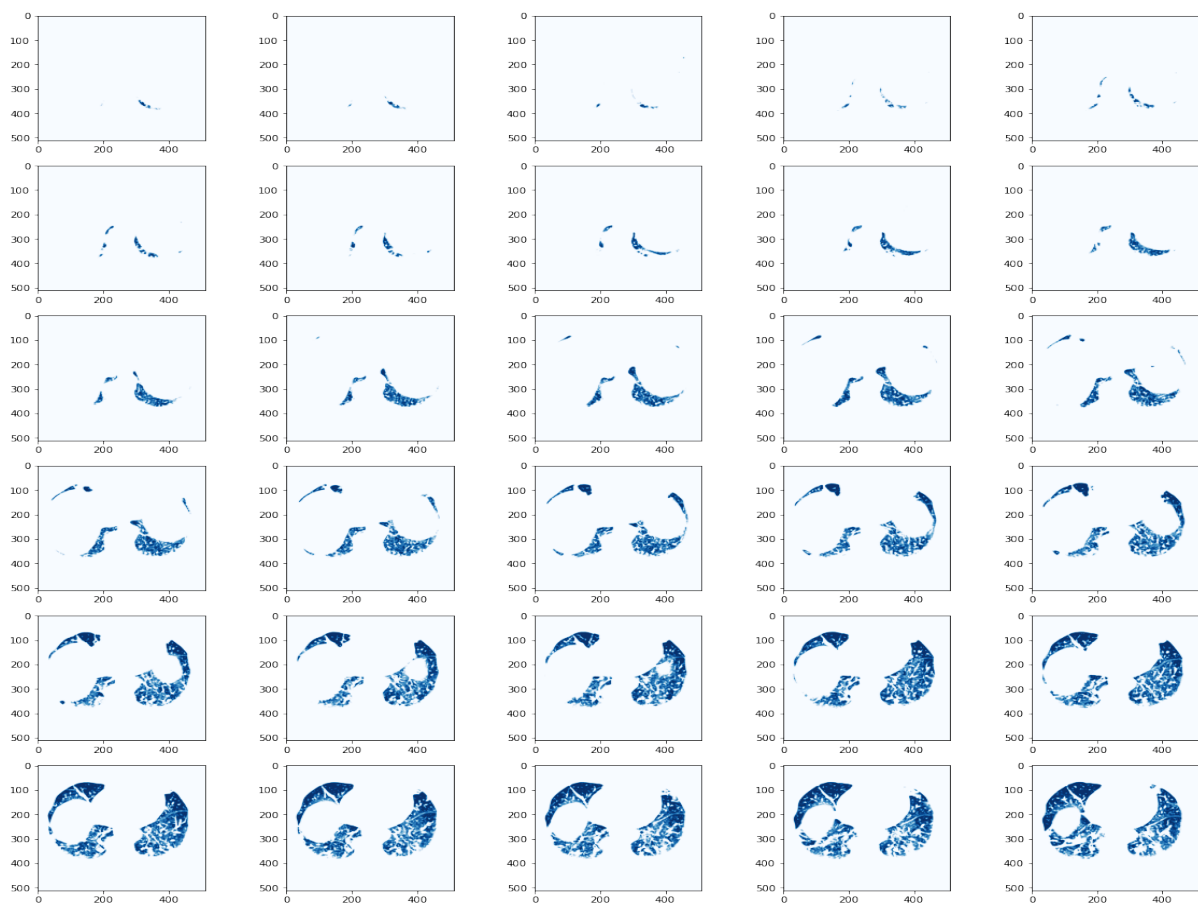


Figure (3.22) Segmented Slices of Lungs

Chapter 4

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.

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.

Transfer learning [5] is a machine learning technique where a model trained on one task is re-purposed on a second related task. It is an optimization that allows rapid progress or improved performance when modeling the second task. Nevertheless, transfer learning is popular in deep learning given the enormous resources required to train deep learning models or the large and challenging datasets on which deep learning models are trained. Transfer learning only works in deep learning if the model features learned from the first task are general. This form of transfer learning used in deep learning is called inductive transfer. This is where the scope of possible models (model bias) is narrowed in a beneficial way by using a model fit on a different but related task.

We will use transfer learning along with quantile regression in our

project. We will discuss briefly about different software and tools, different transfer learning models, quantile regression [2] that we will use in our project and the evaluation metric to evaluate our work in the following section.

4.1 Software and Tools

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

- **Python:** Overall Scripting.
- **Keras**
- **Tensorflow**
- **Numpy**
- **Pandas**
- **Matplotlib:** Data Visualization.
- **Cuda:** Execution.

4.2 Transfer Learning

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).
- DenseNet(e.g. DenseNet121, DenseNet169).
- EfficientNet(e.g. EfficientNetB0-EfficientNetB7).

4.2.1 VGG16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [8]. VGG16 refers to the fact that it has 16 layers that have some weights as shown in 4.1. It has 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.

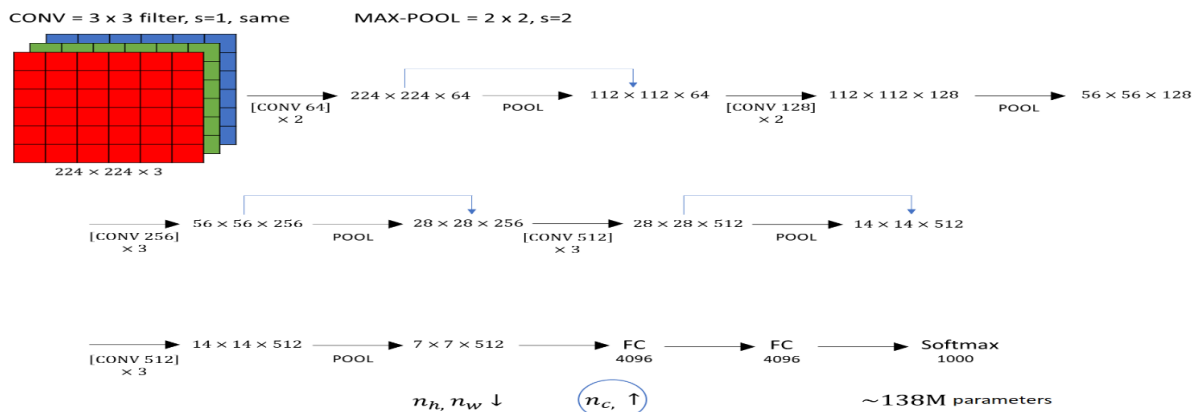


Figure (4.1) VGG16 architecture

4.2.2 VGG19

VGG19 [8] is similar to VGG16. 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 which shown in figure 4.2. 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.

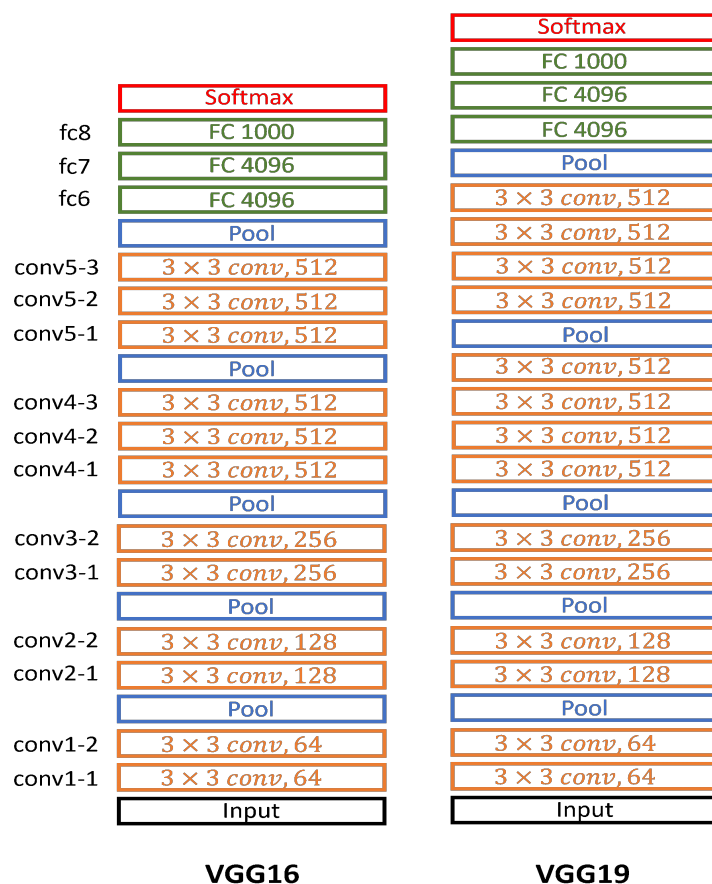


Figure (4.2) Comparison Between VGG16 and VGG19 Architecture

4.2.3 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 which can be found in figure 4.3. The deterioration of training accuracy shows that not all systems are easy to optimize.

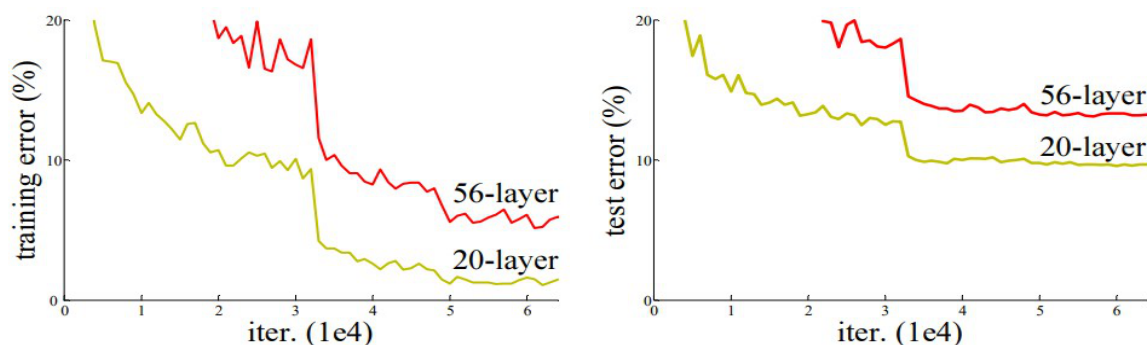


Figure (4.3) Training and testing error of deep network

To overcome this problem, Microsoft introduced a deep residual learning framework [1] shown in 4.4. Instead of hoping every few stacked layers directly fit a desired underlying mapping, they explicitly let these 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.

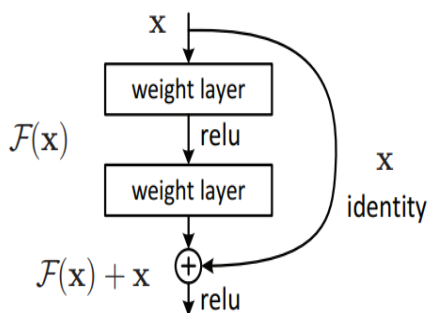


Figure (4.4) The residual block

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.

From figure 4.5, 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.

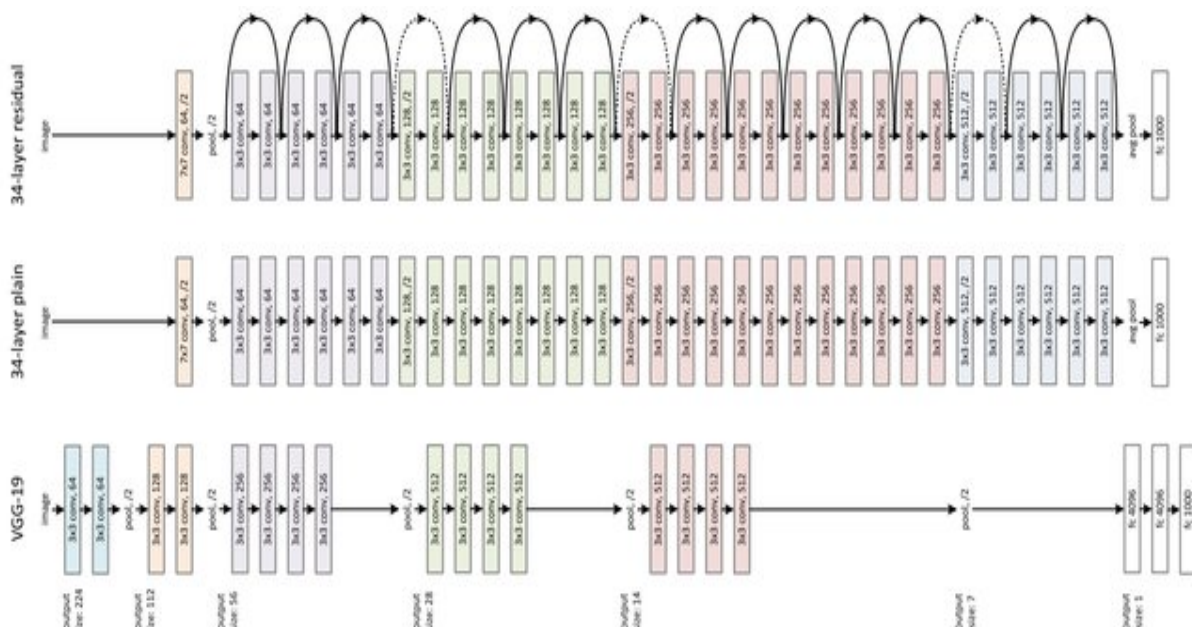


Figure (4.5) ResNet34 architecture in comparison with VGG19

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 1×1 convolutions). 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).

4.2.4 EfficientNet

EfficientNet [9] is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice in 4.6 that arbitrary scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. For example, if we want to use 2^N times more computational resources, then we can simply increase the network depth by α^N , width by β^N , and image size by γ^n , where α, β , are constant coefficients determined by a small grid search on the original small model. EfficientNet uses a compound coefficient ϕ to uniformly scales network width, depth, and resolution in a principled way.

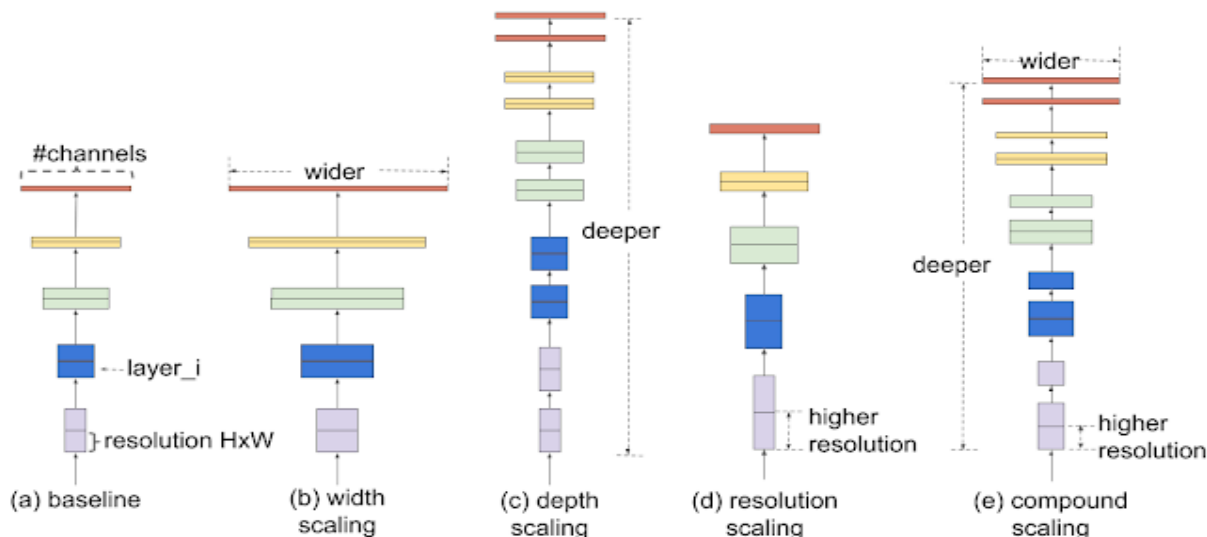


Figure (4.6) Comparison of Different Scaling Methods.

The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. The base EfficientNet-B0 network is based

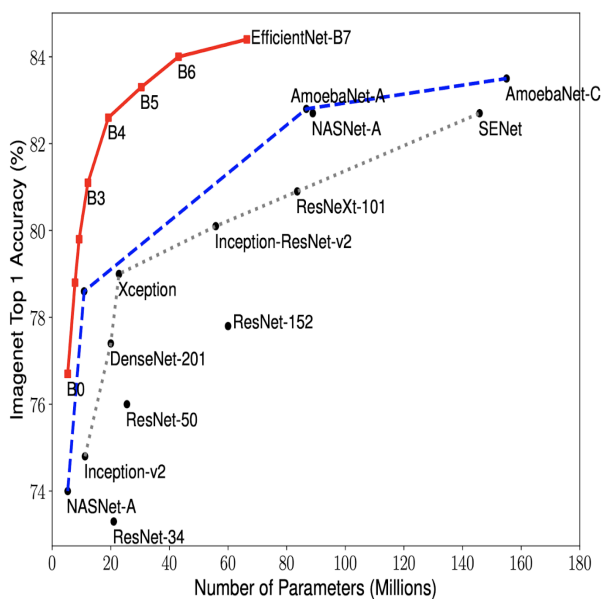


Figure (4.7) Performance of EfficientNet

on the inverted bottleneck residual blocks of MobileNetV2, in addition to squeeze-and-excitation blocks. This compound scaling method consistently improves model accuracy and efficiency for scaling up existing models such as MobileNet (+1.4% imagenet accuracy), and ResNet (+0.7%), compared to conventional scaling methods.

Comparing EfficientNets with other existing CNNs on ImageNet. In general, the EfficientNet models achieve both higher accuracy and better efficiency over existing CNNs, reducing parameter size and FLOPS by an order of magnitude. For example, in the high-accuracy regime, EfficientNet-B7 reaches state-of-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on CPU inference than the previous Gpipe. Compared with the widely used ResNet-50, EfficientNet-B4 uses similar FLOPS, while improving the top-1 accuracy from 76.3% of ResNet-50 to 82.6% (+6.3%).

Figure 4.7 shows the comparison of different model size vs accuracy. EfficientNet-B0 is the baseline network, while Efficient-B1 to B7 are obtained by scaling up the baseline network. In particular, our EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 / 97.1% top-5 accuracy, while being 8.4 \times smaller than the best existing CNN.

Though EfficientNets perform well on ImageNet, to be most useful, they should also transfer to other datasets. To evaluate this, EfficientNets

were tested on eight widely used transfer learning datasets. EfficientNets achieved state-of-the-art accuracy in 5 out of the 8 datasets, such as CIFAR-100 (91.7%) and Flowers (98.8%), with an order of magnitude fewer parameters (up to $21\times$ parameter reduction), suggesting that EfficientNets also transfer well. By providing significant improvements to model efficiency, EfficientNets could potentially serve as a new foundation for future computer vision tasks.

4.3 Quantile Regression

Before going deep in quantile regression, let's look at a few concepts. Quantiles are points in a distribution that relates to the rank order of values in that distribution. The middle value of the sorted sample (middle quantile, 50th percentile) is known as the median.

Regression is a statistical method broadly used in quantitative modeling. Standard linear regression techniques summarize the relationship between a set of regressor/input variables and the outcome variable, based on the conditional mean. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of outcome variables.

Standard linear regression uses the method of least squares to calculate the conditional mean of the outcome variable across different values of the features. Quantile regression is an extension of Standard linear regression, which estimates the conditional median of the outcome variable and can be used when assumptions of linear regression do not meet.

Quantile regression methodology allows understanding relationships between variables outside of the mean of the data, making it useful in understanding outcomes that are non-normally distributed and that have nonlinear relationships with predictor variables. Quantile regression allows the analyst to drop the assumption that variables operate the same at the upper tails of the distribution as at the mean and to identify the factors that are important determinants of variables. Quantile regression can be used in many purpose

- To estimate the median, or the 0.25 quantile, or any quantile.
- Key assumption of linear regression is not satisfied.
- Find Outliers in the data.
- Increase in error variance with increase in outcome variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (4.1)$$

In equation 4.1 p is the number of regressor variables and n is the number of data points. The best linear regression line is found by minimizing the mean square error shown in equation 4.2.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 + (\beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 \quad (4.2)$$

Quantile Regression Model Equation for the τ quantile is

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip} \quad (4.3)$$

In equation 4.3 p is the number of regressor variables and n is the number of data points. The best Quantile regression line is found by minimizing by minimizing median absolute deviation.

$$MAD = \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - (\beta_0(\tau) + (\beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip})))^2 \quad (4.4)$$

Here in equation 4.4 the function ρ is the check function which gives asymmetric weights to the error depending on the quantile and the overall sign of the error. Mathematically, ρ takes the form

$$\rho_\tau(u) = \tau \max(u, 0) + (1 - \tau) \max(-u, 0) \quad (4.5)$$

4.4 Evaluation Metric: Laplace Log Likelihood

We will use a modified version of the laplace log likelihood to evaluate our project. In medical applications, it is useful to evaluate a model's confidence in its decisions. Accordingly, the metric is designed to reflect both the accuracy and certainty of each prediction. For each true FVC

measurement, we predict both an FVC and a confidence measure (which is standard deviation σ). The metric is computed as:

$$\sigma_{clipped} = \max(\sigma, 70), \quad (4.6)$$

$$\Delta = \min(|FVC_{true} - FVC_{predicted}|, 1000), \quad (4.7)$$

$$metric = -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\sigma_{clipped}). \quad (4.8)$$

Chapter 5

Experiments & Results

5.1 Overview

In transfer learning we train the model with a dataset and after we train the same model with another dataset that has a different distribution of classes, or even with other classes than in the training dataset.

In Fine-tuning, another approach of transfer learning, we have a dataset, and we use let's say 80% of it in training. Then we train the same model with the remaining 20%. Usually, we change the learning rate to a smaller one, so it does not have a significant impact on the already adjusted weights.

Training a deep convolutional neural network (CNN) from scratch is difficult because it requires a large amount of labelled training data and a great deal of expertise to ensure proper convergence. A promising alternative is to finetune a CNN that has been pre-trained using, for instance, a large set of labelled natural images. However, the substantial differences between natural and medical images may advise against such knowledge transfer.

We complete this project in two parts. in the first part we will work with image data along with some portion of categorical data to build a model. In the second part we will use the model and to generate new data frame to work with. Using that data frame with some modification and calculation we build the final model to predict the progression.

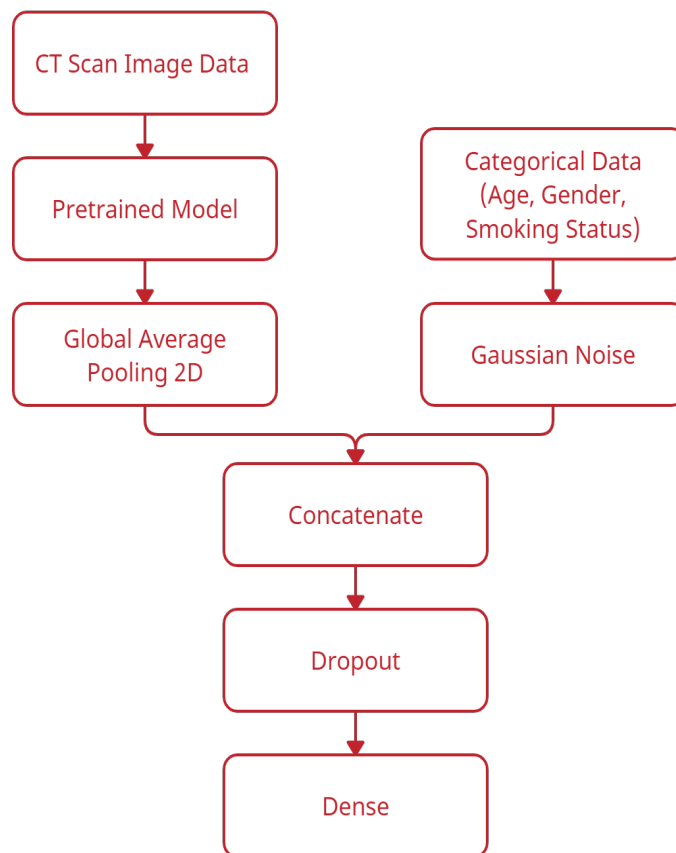


Figure (5.1) Model architecture using image data and partial categorical data

In the first part, we get all the image and prepare it for training. We don't want to train all the image data at once because it will raise out of memory (OMM) error. To avoid the error we create an image generator so that we can train one chunk of data at a time. After that we build the model. The architecture of the model is shown in 5.1. At the beginning we get a model and add input layer of size $512 \times 512 \times 1$ because all models input layer size is 224×224 then we set image data as the input of the pre-trained model after that we add a global average pooling 2D layer. We process the categorical data (age, gender and smoking status) and pass it through a gaussian noise layer. We have done this two thing concurrently. After that we concatenate both of them and after that we add a dropout layer and pass it through a dense layer. Then we proceed for the training phase. At the training phase we use early stopping. We use it because too many epochs can lead to overfitting of the training dataset, whereas too few may result in an underfit model. Early stopping is a method that allows us to specify an arbitrary large number of training epochs and stop training once the model performance stops improving on a hold out

validation dataset. We use reduced learning rate on plateau in order to reduce learning rate when a metric has stopped improving. Models often benefit from reducing the learning rate by a factor of 2-10 once learning stagnates. This callback monitors a quantity and if no improvement is seen for a 'patience' number of epochs, the learning rate is reduced. After the completion of training we plot the training history to visualize how well our model has trained which will be discussed in the next section.

Patient_week	FVC	Confidence
ID00419637202311204720264.-12	3111.484416	161.671271
ID00421637202311550012437.-12	2857.890357	200.935648
ID00422637202311677017371.-12	2007.297131	153.969624
ID00423637202312137826377.-12	3420.942369	206.201271
ID00426637202313170790466.-12	2974.510875	121.335843

Table (5.1) Example submission table

In the second part of our project, We have created two variable for storing train and test set namely 'tr' for train and 'chunk' for test. As we tend to predict the fvc value with respect to week, we have stored the sample submission information into a variable called 'sub'. As you can see from the table 5.1, it has three features called 'Patient_week', 'fvc' and 'confidence' and every entry in the 'patient_week' field is in concatenated form. So we split them into 'patient.id' and 'week' field and created new features in the 'sub' table to store them. In the 'tr', 'chunk', and 'sub'

Patient ID	Start_FVC
ID00007637202177411956430	2315
ID00009637202177434476278	3660
ID00010637202177584971671	3523
ID00011637202177653955184	3326
ID00012637202177665765362	3418

Table (5.2) Example of starting week FVC for each patient

table we have created a feature called 'where' and initialize it with 'train', 'val', and 'test' respectively. Then we merge them together to create one single table called 'data'. In the 'data' table we have created a new feature called 'start_week' and assign 'nan' values (which means not a number) to 'start_week' corresponding to the 'test' rows. After that we have calculated arithmetic mean of weeks for each patient and assign 'start_week' column by that value for each patient. Then we create a new table called 'base'

Patient ID	...	Male	Female	Ex smoker	Never smoked	Currently smokes
ID00419637202311204720264	...	1	0	1	0	0
ID00421637202311550012437	...	1	0	0	0	1
ID00422637202311677017371	...	0	1	0	1	0
ID00423637202312137826377	...	1	0	0	1	0
ID00426637202313170790466	...	0	1	0	1	0

Table (5.3) Updated data table

for only 'start_week' for each patient and add cumulative sum of weeks in each row for each patient to get the starting week fvc value as shown table 5.2. In the 'base' table we have the initial weeks fvc value for each table. We merge them with 'data' table with respect to patient ID. In the data table we have created a new feature called 'base_week'. We subtract 'week' from 'start_week' for each entry and put that value in 'base_week'. In order to save memory space delete 'base' table as we don't need it anymore. We realize that the 'sex' and 'smoking_status' are categorical in nature so we created those features in the table and convert them in

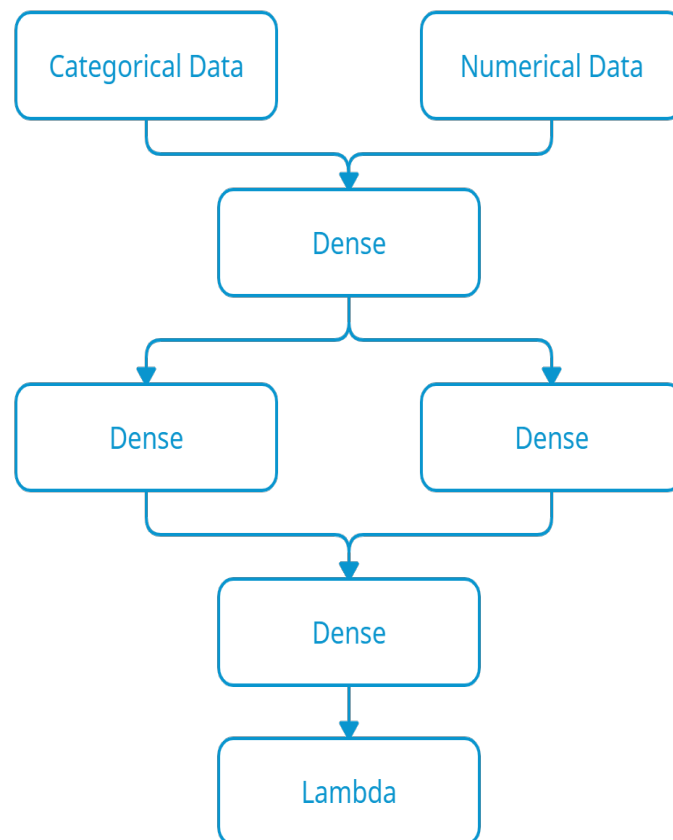


Figure (5.2) Final model architecture

boolean value. As you can see from 5.3 the table below if the patient is a male the corresponding entry is 1 otherwise zero and same for smoking status. Now, we have normalize all the value of 'age', 'fvc', 'week', 'percent' for each patient and add them in the 'feature' table. In order to simplify the calculation we did all these analysis in a single dataframe. After doing that we again split them in train, validation and test set. Then we have build another model to train the new dataset which is shown in 5.2. For building the model we took the processed categorical and numerical data and pass it through multiple dense layer. After that we have trained the model and validate it with validation data. All the information about training, validation, testing are discussed in the next section.

5.2 Result

First we trained our neural network from scratch. In this case we make a dictionary consist of several model including 8 different type of Efficientnet model, ResNet50 and VGG16. At this point we use only the architecture of different model and train them from scratch without using the pretrained weight. We get a model and add input layer of size $512 \times 512 \times 1$ because all models input layer size is 224×224 . After that we added a pooling layer at the end of the model architecture. We use Dropout layer. The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting.

We use batch size of 8, choose 30 for epoch and 0.003 as learning rate. We used ModelCheckpoint in order to save the best model. The best model is the one which has the lowest validation loss. We also used Reduced on platow to reduce the learning rate when validation loss has stopped improving. If improvement stops for consecutive 5 epochs learning rate will be reduced by half($\frac{1}{2}$).

We have trained 6 different model, which include 4 different type of Efficientnet namely B0, B1, B2, B4. And ResNet50 and VGG16. All these models are different in terms of parameters and architecture. We have trained all these models with the same hyperparameters so that we can compare all of them. EfficientNet B0 has the lowest validation loss of 3.724704 and at that epoch train loss was 4.451072. For EfficientNet

B1, lowest validation loss was 3.287487 and train loss was 4.078069. For ResNet50, lowest validation loss was 1.505803 and train loss was 4.399539. As we can see in the table 5.4, the train loss is almost the same for all the 6 model. In terms of validation loss it stays almost the same for four different EfficientNet model. But for ResNet50 and VGG16 the validation loss is slightly lower. But when we test the model we find that ResNet50 and VGG16 is overfitting. Our assumption is that ResNet50 and VGG16 are deep in architecture which require a huge amount of training data but our dataset is not that kind of big, this may led to overfit the data. Figure 5.3 shows the training and validation accuracy over 30 epoch for all six model. But if we take a look a each graph of figure 5.3 we cannot see any changes in training loss. This is because training loss is very small relative to validation loss. To mitigate the problem we have plot the smooth curve of training loss and validation of these model shown in figure 5.4 and 5.5 respectively.

We have evaluated our trained model on a modified version of the Laplace Log Likelihood discussed in 4.8. In medical applications, it is useful to evaluate a model’s confidence in its decisions. The error is thresholded at 1000 ml to avoid large errors adversely penalizing results, while the confidence values are clipped at 70 ml to reflect the approximate measurement uncertainty in FVC. The final score is calculated by averaging the metric across all test set Patient_Weeks (three per patient). Note that metric values will be negative and higher is better. We defined the laplace log likelihood function in order to use it with some examples and understand the metric. We find that -8.023 is the default score to beat while cross-validating models on train data. Any model scoring worse than this (which means lower than this) is not useful.

Model Name	Train Loss	Validation Loss
EfficientNet B0	4.451072	3.724704
EfficientNet B1	4.078069	3.287487
EfficientNet B2	4.495673	3.854002
EfficientNet B4	4.185572	3.281747
ResNet 50	4.399539	1.505803
VGG16	4.187896	1.489869

Table (5.4) Comparison of Training Loss And Validation Loss with Different Model

In this phase we have trained the model for 600 epoch and validate it with 5 fold cross validation. Speaking about validation, we make it as close as possible to the scoring method used by organizers of the kaggle competition. Originally, they scored only 3 last predictions (last 3 weeks for each patient), so we developed a similar validation framework. For the test set, we included only patients not present in the training set and used only 3 last weeks for scoring. We have achieved the evaluation score as high as to -6.6767.

In table 5.5 we compare our result with the result found in the paper [4] and also with the Kaggle public leaderboard. We have sorted the comparison table in descending order so we can see which score is higher. As you can see we have achieved better result than others.

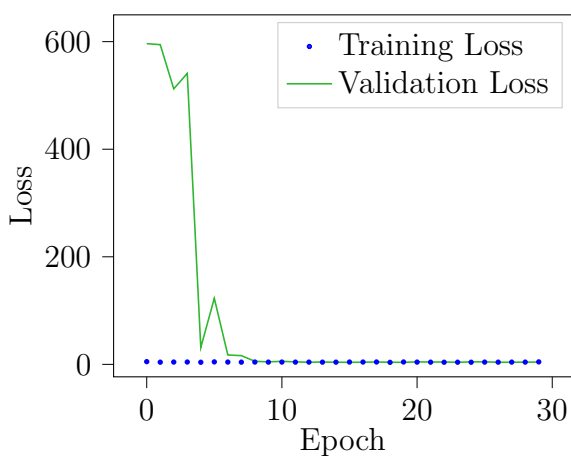
Model Name	Score
Our Model (Efficientnet B1 + Quantile Regression)	-6.6767
Elastic Net Regression (paper [4])	-6.73
Kaggle Leaderboard position 1	-6.7465
Kaggle Leaderboard position 1	-6.7471
Kaggle Leaderboard position 1	-6.7480
Ridge Regression	-6.81
Multiple Quantile Regression	-6.92

Table (5.5) Comparison of score of different models

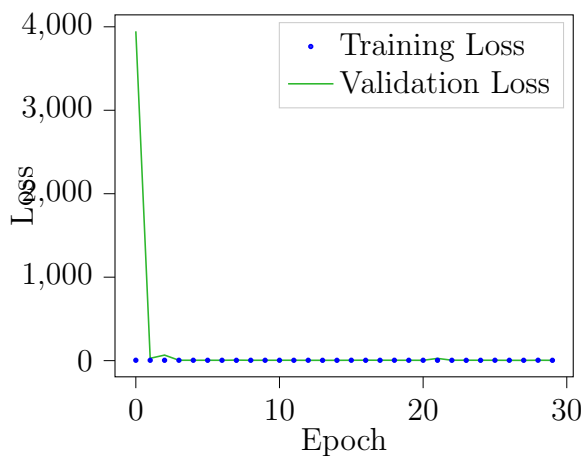
As we discuss earlier, our intention of this project is to predict the FVC value along with a confidence value. Table 5.6 shows the sample of our prediction.

Patient_week	FVC	Confidence
ID00419637202311204720264.-10	2922.207077	214.653266
ID00419637202311204720264.-11	2924.974276	214.762988
ID00419637202311204720264.-12	2927.741566	214.945554

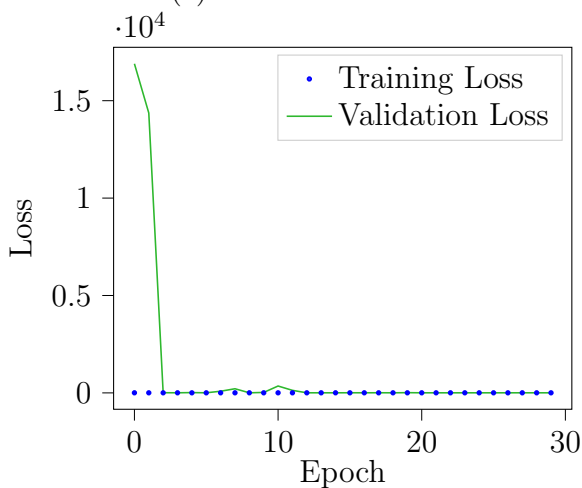
Table (5.6) Example submission table



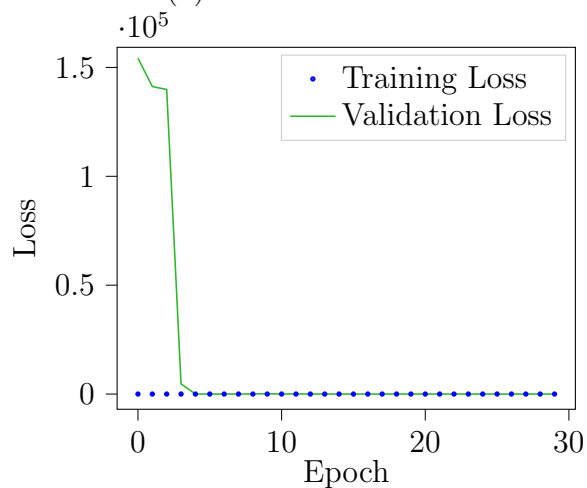
(a) Efficientnet B0



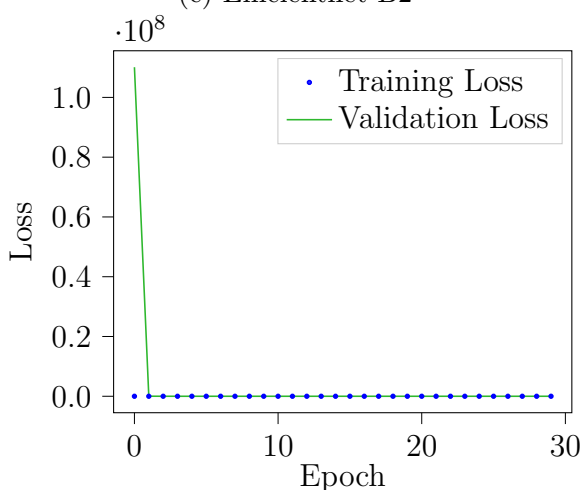
(b) Efficientnet B1



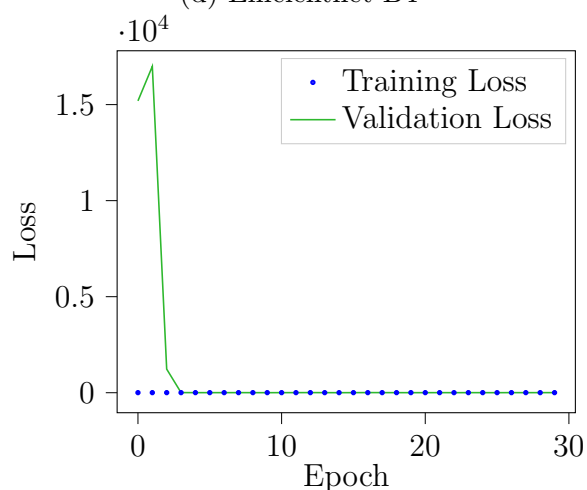
(c) Efficientnet B2



(d) Efficientnet B4



(e) ResNet50



(f) VGG16

Figure (5.3) Training & Validation Loss

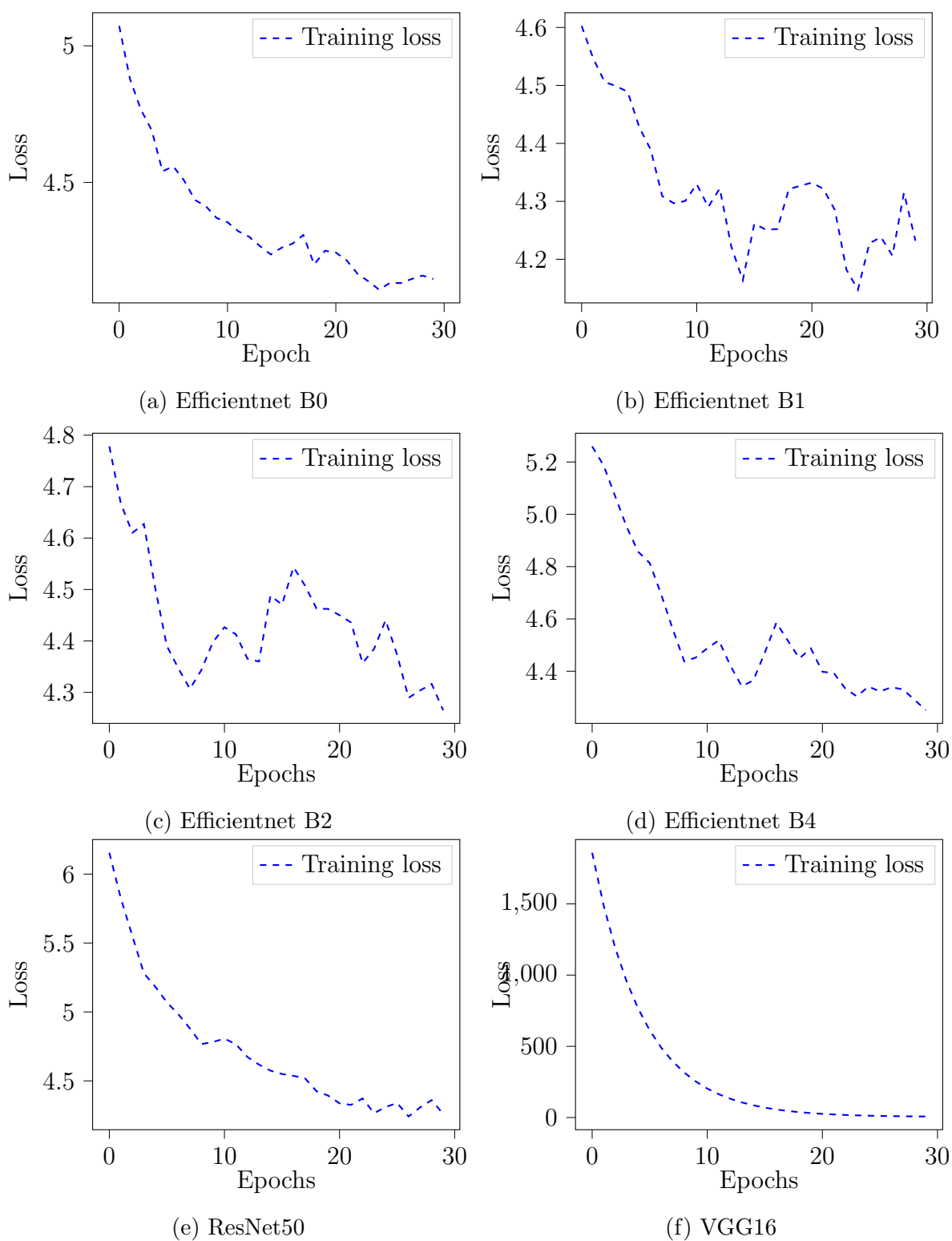
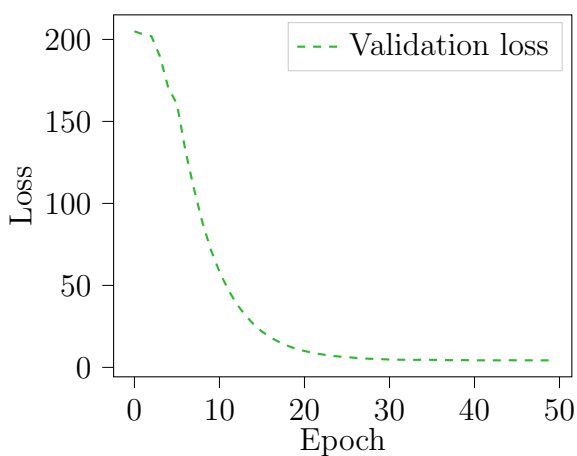
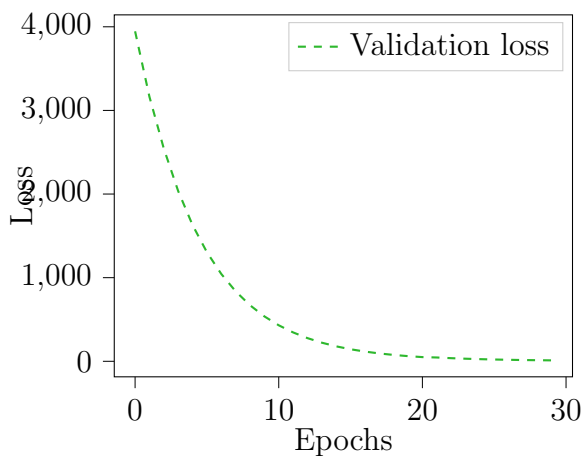


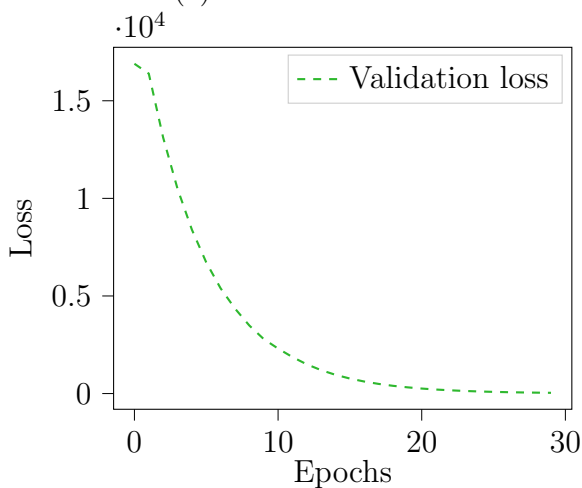
Figure (5.4) Training loss curve after smoothing



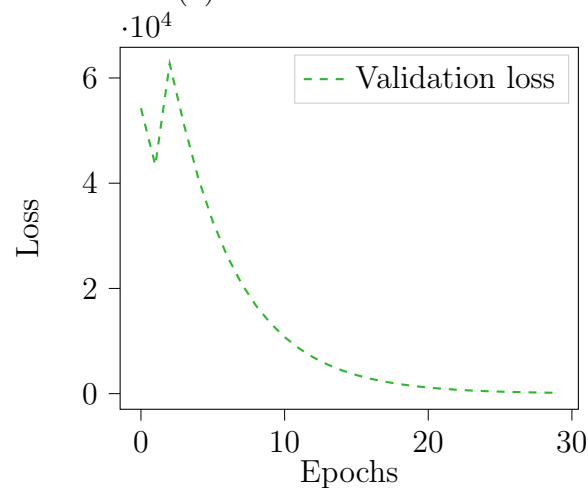
(a) Efficientnet B0



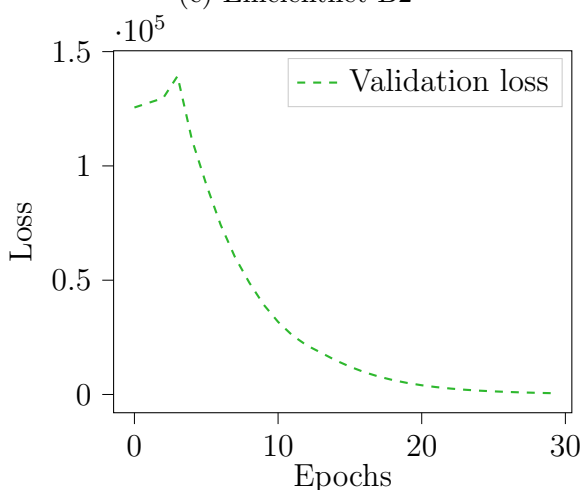
(b) Efficientnet B1



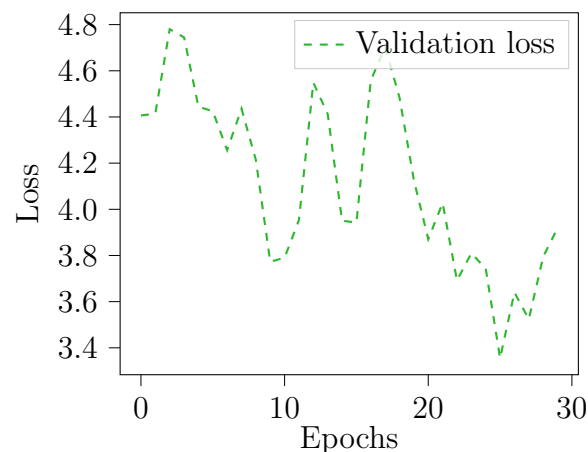
(c) Efficientnet B2



(d) Efficientnet B4



(e) ResNet50



(f) VGG16

Figure (5.5) Validation loss curve after smoothing

Chapter 6

Discussion

In this chapter, we are going to explain about the limitations or challenges related that we have faced to complete our project and how we can make our model more efficient by doing further developments in the future.

6.1 Future Work

We would like to work with this project further. Some of our future development goals are as follows:

- We would like to build a model from scratch and train it without using transfer learning in order to see if transfer learning works better or not.
- We would like to use Auto Encoder Decoder model for this problem.
- We will fine tune our existing model by tuning hyper parameters.

6.2 Challenges Limitations

While working on the project we have faced many limitations and challenges. Some of them are as follows:

- Lack of resources. There is not much resources available to conduct

the research because deep learning approaches is new to the medical field.

- Lack of computational power. At first we have struggled to train our model because it takes 14-15 hours to train a model while epoch was set to 25. We had to struggle with it for more than four month.
- Conflict between different libraries while installing.

6.3 Conclusion

In our work we have shown how we can use deep learning technique to predict the idiopathic pulmonary fibrosis progression. This type of prediction analysis, if applied effectively in medical sectors, can help patients by analysing their lungs condition from Ct scan and the corresponding other information so that treatment can be started as early as possible. Early treatment can improve the survival rate of a patient. Thus, deep learning can help medical practitioners to understand their prognosis in a better way when they are first diagnosed with idiopathic pulmonary fibrosis. Hence, deep learning algorithms are adding significant value to the healthcare industry as well as to the society aiming towards a healthy and normal social lifestyle.

References

- [1] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [2] Roger Koenker and Kevin F Hallock. “Quantile regression”. In: *Journal of economic perspectives* 15.4 (2001), pp. 143–156.
- [3] Hyunkwang Lee et al. “An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets”. In: *Nature Biomedical Engineering* 3.3 (2019), p. 173.
- [4] Sampurna Mandal et al. “Prediction Analysis of Idiopathic Pulmonary Fibrosis Progression from OSIC Dataset”. In: *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE. 2020, pp. 861–865.
- [5] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [6] Margaret L Salisbury et al. “Idiopathic pulmonary fibrosis: gender-age-physiology index stage for predicting future lung function decline”. In: *Chest* 149.2 (2016), pp. 491–498.
- [7] Yu Shi et al. “Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach”. In: *Artificial intelligence in medicine* 100 (2019), p. 101709.
- [8] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [9] Mingxing Tan and Quoc V Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *arXiv preprint arXiv:1905.11946* (2019).
- [10] Ana Adriana Trusculescu et al. “Deep learning in interstitial lung disease—how long until daily practice”. In: *European Radiology* (2020), pp. 1–8.
- [11] Shudong Wang et al. “Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy”. In: *Open Medicine* 15.1 (2020), pp. 190–197.
- [12] Yutong Xie et al. “Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT”. In: *IEEE transactions on medical imaging* 38.4 (2018), pp. 991–1004.