**Software Requirement Specification**

* 1. **Purpose:**

To analyze the lung function decline in patients with Pulmonary Fibrosis (PF) using deep learning techniques. PF is a disorder with no known cause and no known cure, so analyzing the patient’s prognosis and detecting the decline in patient’s lung function is very much needful. The project aims to predict the forced vital capacity of the patient based on which he/she can choose further treatment.

* 1. **Abbreviations:**

ML – Machine Learning

FVC – Forced vital capacity

XGB – Extreme Gradient Boosting

API – Application programming interface

DICOM – Digital Imaging and Communications in Medicine

**1.3 Overview:**

The project is aimed at determining the patient’s FVC value by analyzing the lungs CT scan images taken at the week 0. The dataset contains FVC values taken both before and after the week 0, which are negative values and positive values respectively. The dataset also contains percentage in lung function decline and their smoking status. The data is visualized using various tools and techniques such as pandas, matplotlib and seaborn. We have extracted the features from 2D slices of 3D image of patient’s CT scan using ResNet50. The feature vectors along with the tabular data are then fitted into XGB Regressor and predicted the FVC value. Currently model is deployed on local host using Flask API.

**List of Functions:**

|  |  |
| --- | --- |
| Data Visualization | Analyzing and finding patterns in data |
| Data Preprocessing | Treatment of inconsistencies in the data |
| Feature Extraction | Extracting the features from CT scan image |
| Model Building | Creating ML models to fit the data |
| Model Selection | Selecting the best model |
| Model Deployment | Deploying the model for patients use |

**2. Functional Description:**

* **Data visualization –** The data needs to be presented in a visually appealing way using graphs, charts, and plots to make it easier for users to understand and draw insights from data. We have used pandas, matplotlib and seaborn find the patterns of both tabular data and CT scan images.
* **Data preprocessing –** Looking deeper into the data to treat null and duplicate values is necessary. Later converted the categorical values into numerical values using Label Encoding. The CT scan images were given in dicom format, so converted it into jpg format for easy fitting into the ML model. Also, identified the patients whose FVC values of week 0 were given and collected their CT scan images.
* **Feature extraction –** The next step was extraction of important features from the CT scan images using ResNet50 model and added those values to the tabular data to fit into the model. Used autoencoder for dimension reduction of image features. Also extracted the metadata from the CT scan images.
* **Model building –** The modified tabular data is fitted into XGB Regressor and predicted the corresponding FVC values and calculated the accuracy.
* **Model selection –** Performed the evaluation and selected the best model which give more accuracy.
* **Model deployment –** The model is now ready for the deployment and we deployed it in local host. By uploading the lung’s CT scan image, the model correctly predicts its FVC value.

**3. Specific Requirements**

**3.1 Software interfaces:**

* Python
* Pandas
* Matplotlib
* Seaborn
* Numpy
* Sklearn
* Pydicom
* Tensorfow
* Skimage
* XGB Regressor

**3.2 Hardware interfaces:**

* 8GB RAM
* 512GB HDD
* Intel core i5 processor
* Intel Graphics Card