# **RESEARCH METHADOLOGY**

## **3.1 INTRODUCTION**

This chapter explains in detail the methods and materials to be used for the success of the research. The research methodology would be explained in details i.e. the algorithms, data source and development of model.

## **3.2 RESEARCH DESIGN**

The research was done in the following steps, with each step thought out critically, in order to achieve the stated objective:

* Date Collection
* Data Processing/ Analysis
* Feature Selection
* Classification model
* Implementation/ Development of Predictive Model
* Performance Evaluation and Optimization

## **3.3 DATA COLLECTION**

When building any model, the presence of data is essential as data can be used to both validate and build the model (in the case of statistical models). The same applies in the development of machine learning model, data is what the machine learns from and also make new deductions from, therefore it was of huge important to secure access to data.

In order to develop a model that would help make Lung Cancer Prediction, data on various patient was needed. The data would typically show the activity of the patient that affected the lungs and the overall effect they had on the lungs. A data that fit this requirement was acquired from the website data.world. The data consist of Cross-sectional collection of 1000 subjects of both genders, aged between 14 to 76. Each subject was put on several test in other know their levels in some of the activities that are in direct relationship with the condition of the lungs, for example Smoking, Fatigue, Alcohol, etc. amongst the subjects are those with confirmed cases of Lung Cancer and also those without lung cancer.

In line with ***ethical methods***, the name Names of Patients removed and replaced with Index Tag. Also, clinical Settings showing clinicians, Hospitals, Clinics and Locations names are removed and referenced with clinical settings tags.

Table 3.1 Features Information

|  |  |
| --- | --- |
| Header | Feature Information |
| Patient ID | Subject Identification Number |
| Age | Age of Subjects in Years |
| Gender | The Gender of Subjects Male or Female) |
| Air Pollution | Subject Exposure to Air Pollution |
| Alcohol Use | How Often the Subjects Uses Alcohol |
| Dust Allergy | Allergic Reaction to Dust Rating |
| Occupational Hazards | Exposure to occupational hazards that effect Lungs |
| Genetic Risk | Inheritance of Diseases from Parents (Genetics) |
| Chronic Lung Disease | Level of Lung disease in Subject |
| Balanced Diet | Eating Habit of Subjects |
| Obesity | Body size of Subject |
| Smoking | Frequency of Smoking |
| Passive Smoker | Frequency of Inhalation of Smoke, SHS (Secondhand Smoking) |
| Chest Pain | Presence of chest Pain in Subject |
| Coughing of Blood | Frequency of Coughing out Blood |
| Fatigue | How often Subject Feel Fatigued |
| Weight Loss | Severity of wight Loss |
| Shortness of Breath | Frequency of Shortness of Breath |
| Wheezing | Subjects that Wheeze’s |
| Swallowing Difficulty | How often and How painful a Subject Swallow |
| Clubbing of Finger Nails | Severity of Nail clubbing |
| Frequent Cold | How often Subjects get ‘Cold’ |
| Dry Cough | Severity of Dry cough |
| Snoring | Snoring Noise Level |
| Level | The rating on Level of Lung Cancer (Low, Medium, High) |

## **3.4 DATA PROCESSING/ ANALYSIS**

Data Processing also known as Data Cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. It also refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data (Wu, 2013).

To perform the cleaning of the data, four modules in **Python** were used, they include Pandas, NumPy, Matplotlib, and Seaborn. The following steps were then taken:

1. The first step was to understand the data, and the following analysis was done to the parts of the data that are numerical:

* Counted the number of entries or values
* Calculated the mean of the values
* Calculated standard deviation of each features value.
* Identified the min and the max value in the data and the corresponding points in between.

1. Exploratory data analysis and cleaning of Age feature: To understand the feature better,

* A bar chart showing the relationship Age vs Risk of Cancer (Level).
* A grouped bar chart showing the relationship between Age and the various Level of Lung Cancer (Risk) i.e. Low, Medium and High.
* Dived the age group into 5 groups and called it Age bands, this helped section the age into similar age groups.
* A group chart was then plotted to show the relationship between the age groups and Lung Cancer Risk.

1. Exploratory data analysis and cleaning of Gender feature:

* A group chart showing relationship between Gender and Lung Cancer Risk
* The gender values (Male and Female) were the encoded to 1 and 0. With 0 being Male and 1 being Female.

1. A group charts showing relationship between;

* Alcohol and Lung Cancer Risk.
* Dust Allergy and Lung Cancer Risk.
* Air Pollution and Kung Cancer Risk.
* Genetic Risk Feature and Lung Cancer Risk.
* Chronic Lung Disease and Lung Cancer Risk.
* Occupational Hazard and Chronic Lung Disease.
* Balanced Diet and Lung Cancer Risk.
* Obesity and Lung Cancer Risk
* Obesity and Balanced Diet
* Smoking and Lung Cancer Diet
* Passive Smoker and Lung Cancer Diet
* Chest Pain and Lung Cancer Risk
* Chest Pain and Smoking
* Chest Pain and Occupational Hazard
* Coughing of Blood and Lung Cancer Risk
* Cleaning of Fatigue Feature and Lung Cancer Risk
* Weight Loss and Lung Cancer Risk
* Shortness of Breath and Lung Cancer Risk
* Wheezing and Lung Cancer Risk
* Difficulty Swallowing and Lung Cancer Risk
* Frequent Cold and Lung Cancer Risk
* Dry Cough and Lung Cancer Risk
* Snoring and Lung Cancer Risk

All of this charts were drawn in other to understand the relationship between the features and the target features( Lung Cancer Risk(Level))

1. The next step was to perform correlation analysis, in order to identify features that are key. A correlation plot was then plotted to find the correlation between the many features of the data.

## **3.5 FEATURE SELECTION**

To be decided on

## **3.6 CLASSIFICATION MODEL**

The model will be used in this research is neural network algorithm with quasi-newton method for optimization. Artificial Neural Network (ANN) is a mathematical model, which can solve the problem of prediction. (Nasser & Abu-Naser, 2019). Neural networks involve input and output layers, as well as (in most cases) hidden layers that transform the input into something so the output layer can use. Each neuron in a layer is connected with each neuron in the next layer through a weighted connection. The neurons in the input layer receive the data and transfer them to neurons in the first hidden layer through the weighted links. Here, the data are mathematically processed and the result is transferred to the neurons in the next layer. Ultimately, the neurons in the last layer provide the network’s output.

### **3.6.1 How Neural Network Model Works?**

Fundamentally, each neuron in a neural network is just a mathematical function. Each neuron computes a weighted sum of its inputs—the larger an input's weight, the more that input affects the neuron's output. This weighted sum is then fed into a non-linear function called an activation function—a step that enables neural networks to model complex non-linear phenomena.

### **3.6.2 Quasi Newton Method as an Optimizer**

The most well-known minimization technique for unconstrained problems is Newton’s Method. It is effective, robust and quadratically convergent. In each iteration, the step update is: . The second derivative need to be calculated analytically and supplied to the algorithm by the user.

However, Newton’s method has the disadvantage of being computationally expensive. The inverse of the Hessian has to be calculated in every iteration, and that is rather costly. Moreover, in some applications, the second derivatives may be unavailable. One fix to the problem is to use a finite difference approximation to the Hessian. The other fix, which is more widely used, is quasi- Newton Methods, where approximate Hessian or inverse Hessian updates are updated in each iteration, while the gradients are supplied.

The general structure of quasi-Newton method can be summarized as follows

Given x0 and initial Hessian approximation B0(inverse Hessian approximation H0);

Evaluate gradient gk.

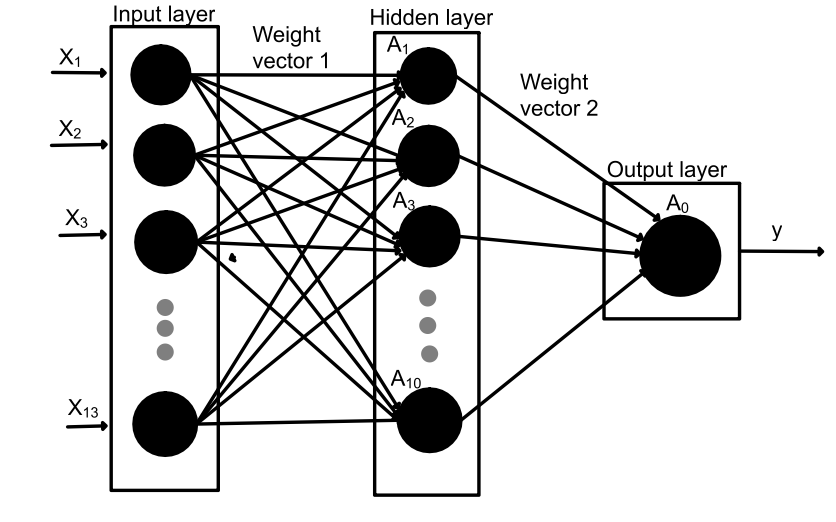
Calculate the update vector sk by line search or trust region methods.

Update Bk+1 or Hk+1 according to the quasi-Newton formulas.

The basic requirement for the updating formula is that the secant condition is satisfied in each iteration, i.e., Bk+1sk = yk

### **Design of Neural Network Model**

Architecture of the purposed model has 3 layers which are one input layer, one hidden layer and one output layer.



### **Selection of Weights**

Initially the weights on all the interconnections are set to be small random numbers between -1 and +1. Then the network will be presented with a trained data set and weights are adjusted in such a way that each weight adjustment increases the likelihood that the network will compute. To adjust weights is calculated and the weights are then changed such that the error decreases.

### **Activation Function**

* Hyperbolic Tangent Function (tanh)

It will be used in hidden layer

* Sigmoid Function

It will be used in output layer to determine whether the patient may have cancer or not.

Tools and Platforms

There are different kinds of libraries and frameworks which are suitable for machine learning. For this research Python Machine Learning libraries like Pytorch, Scikit Learn, Pandas, and NumPy were used in order to build and evaluate the model. The python version used was Python 3.7.