# LILAC : ANANLYZING LIFESTYLE AS RISK FACTOR FOR LUNG CANCER

**Amrita Aash**
Dept. of CSE
IIITD, Delhi
amrita22011@iiitd.ac.in

**Medha**
Dept. of CSE
IIITD, Delhi
medha22110@iiitd.ac.in

**Mohit Gupta**
Dept. of CSE
IIITD, Delhi
mohit22112@iiitd.ac.in

**Ritisha Gupta**
Dept. of CSE
IIITD, Delhi
ritisha22056@iiitd.ac.in

## Abstract

Over the last few years, the modern-day lifestyle has evolved with an increasing count of people adopting sedentary living with continuous exposure to consistent stress and degrading environmental conditions. In such times, early detection of a major risk to develop a highly dreadful disease for timely medication may prove to be pivotal for Bio-Medicine Computational Sciences for improving human health. LILAC provides an automated solution to visualize, analyze, process, and hence predict the expected risk of Lung Cancer through available data records concerning the lifestyle of people. The project attempts to propose a detailed comparison of popular Machine Learning techniques with an aim to predict the risk of Lung Cancer effectively by analyzing and refining the set of available attributes through several feature selection processes to select the factors which strongly affect the chances of Lung Cancer. An effective summation and overall comparison of the executed techniques to select the best technique is a major highlight of the presented work. The best feature selection technique is attributed to Sequential Feature Selection as it led to a significant improvement in the prediction of the risk through relatively easier models with high accuracy. The proposed work enhances the accuracy of models which are relatively easy to build and explain and hence improves the time complexity when the model is built with a huge amount of data available.

***Keywords*** Machine Learning · Feature Selection · Lung Cancer · Prediction System · SVC · Naive Bayes · Decision Trees · Logistic Regression · Performance Analysis

## 1 Introduction

An enormous increase in the proportion of fatalities by cancer drives an immediate need to diagnose, detect and cure a mere lethal infection at an initial stage which might inhibit the human body to eventually proving itself deadly. The observation is derived from the fact that lung cancer is an intrinsic cause of mortality worldwide, and an approach to deal with the same is an effort that should be taken into consideration on a global scale. The latest report by WHO supports the estimation that approximately one million people die from this disease every year, with a higher probability of a patient being a breast, lung, colon, rectum, and prostate cancer patient, amongst

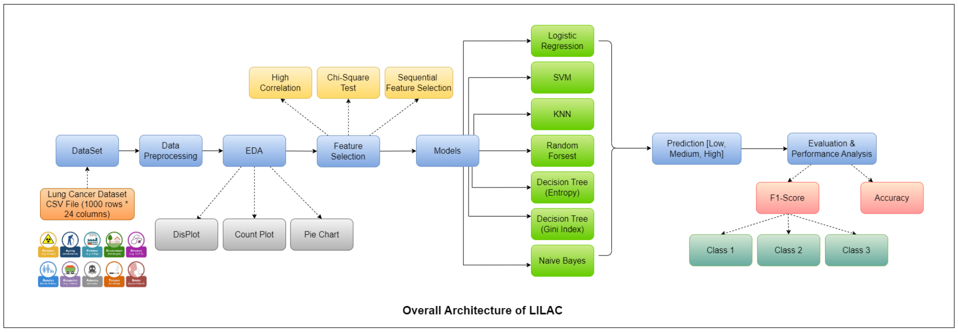


Figure 1: **The system architecture for LILAC. A cumulative review of different Machine Learning and Feature Selection techniques to produce an appropriate prediction with maximum accuracy for low, medium, or high risk of cancer on the basis of lifestyle, health, and environment of an individual**

all other cancers. The number is expected to remain the same or rather increase for the foreseeable future. Early cancer detection can be a very crucial aspect of effective cancer therapy. Several lives will be emancipated if lung cancer is detected at its early stages.

According to the Indian Council of Medical Research (ICMR), lung cancer accounts for $6.9\%$ of deaths occurring due to cancer. As of November 2022, currently, India has approximately 70,275 cases, and it is expected to get doubled by 2025. Numerous findings suggest that a negative lifestyle can be a major factor in determining health outcomes. In this generation, negative habits like smoking, alcohol usage, prepacked food, air pollution, genetic risks, and chronic lung disease can increase the risk of lung cancer, which may prove to be fatal. It is feasible to treat and cure lung cancer if predicted earlier. Cancers have a significant and evident symptomatic presentation in general, but in many cases, the patient remains asymptomatic in its early phases, thus making it extremely inconclusive to ascertain. Predominantly, poor lifestyle and hence early indicative symptoms of cancer are ignored, possibly due to poor health literacy, raised financial costs of hospitals, and the fear of cancer diagnosis. By the time symptoms become evident and patients seek medical help, it reaches out for available clinical treatments. As time passes, it becomes difficult to cure, and hence using this prediction system in conjunction with the doctor's advice can result in concrete decisions, and is possible to initiate treatment earlier.

## 2 Literature Review

Machine Learning is a subset of Artificial Intelligence(AI) that authorizes general computational systems to self-learn from the available data and apply the adapted learning to solve tasks without human intervention to aid and boost human predictions for relevant and related unforeseen data. A solver developed is defined as a trained model that learns to perform a task based on the information provided to it. These solvers perform sophisticated tasks and learn from previous errors to improve their future performance. The most recent research in the domain of Medical Science involves AI and Machine Learning. Next, we present the descriptions of relevant research which has performed lung cancer detection or prediction using different machine-learning models. In the research paper [1], the author has explained that Lung Cancer prediction is concluded using various machine learning classification techniques like Logistic Regression, Decision Tree, SVM, and Naive Bayes. Finally, he infers the fact that the SVM technique achieves the highest accuracy among all the machine learning models. The research paper [2] was aided by the oncologist and hence involved professional intervention for relevant feature selection. The performance of the partitioning-based algorithms was analyzed using only three selected attributes from the total number of attributes of the input data set. The k-means algorithm is efficient for lung cancer data sets with our format. In the research paper [3], different results are produced for each classifier on the lung cancer data set obtained. The classifiers such as KNN, SVM, and Logistic Regression were implemented, and corresponding accuracy rates were obtained. Support Vector Machine has the highest accuracy, with $99.3\%$. The proposed method was applied to the medical data set, which helped doctors to make more correct decisions. In the following research, [4], the author not only builds different models but

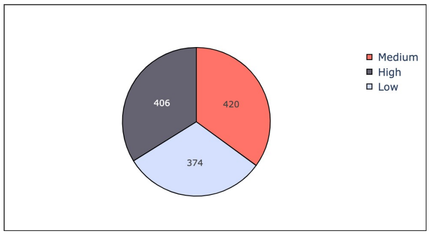
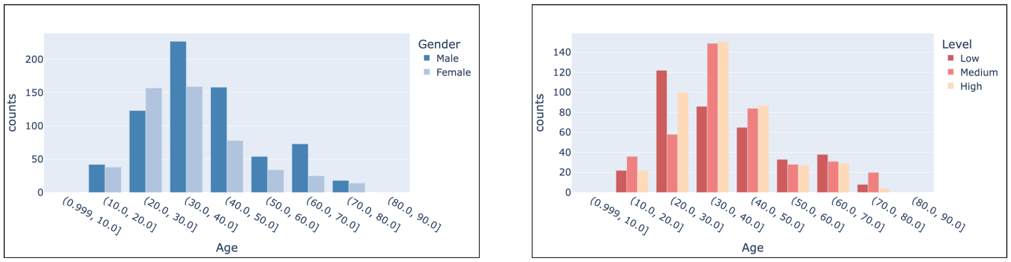


Figure 2: **Pie-Chart for visualization of data samples in three classes as Low, Medium, and High risk of Lung Cancer**



(a)

(b)

Figure 3: **Plots to visualize the data distribution across categories for continuous feature values as 'Age'**

also emphasizes more on the precision, accuracy, or correctness of the model. Therefore, it can be concluded that the best working algorithm for a data set may not be the best for another data set.

## 3 Data Visualisation and Feature Extraction

### 3.1 Data-set Description

The dataset used in this project is publically accessible data. It consists of 1000 samples and 25 features in CSV format. It contains data related to the lifestyle, general health, environment, and habits of hundreds of cancer patients to help predict the risk of Lung Cancer. The dataset contains the following features: Patient ID, Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazard, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Smoking, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing difficulty, Clubbing of Fingers, Frequent Cold, Dry Cough, Snoring and Level where Level is the target variable with three categorical values: Low, Medium and High. Out of the independent features, the columns, Patient Id and Age are numerical whereas the column - Gender is categorical with two unique values of 1 and 2, while the rest are categorical with 8 to 9 classes denoting the severity level of each feature.

### 3.2 Exploratory Data Analysis

As shown in Figure 2, the distribution of three levels of risk of Lung Cancer is balanced across the classes. A balanced data set ensures generous attention toward training the classifier for every possible test outcome and eliminates bias. It is observed that the data is balanced with the following proportions as 33.8%, 35%, and 31.2% of the patients have a high, medium, and low risk of developing Lung Cancer respectively.

From the next Figure 3a, the general inferences are as that the data set contains people from ages 10 to 80, the data set constitutes a major count of people from the age group 20 to 50, and in the age group of 20-30, the count for females is relatively higher than that of males whereas for the rest of all of such buckets, males have has a higher population count. From the adjacent Figure 3b, it can be concluded strongly that the age of a person can not be used as a strong support to predict Lung

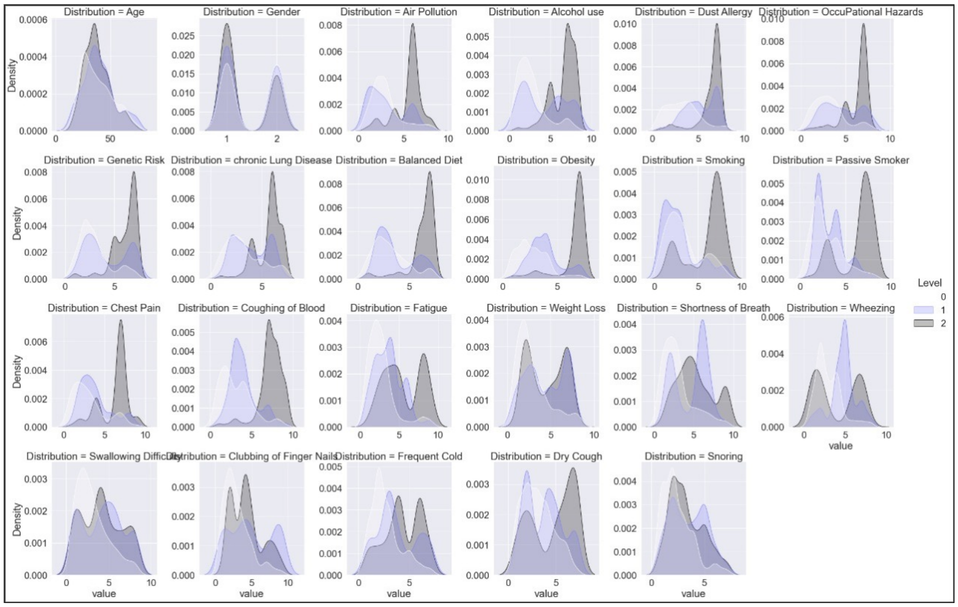


Figure 4: **Distribution Plot for visualizing the Data Distribution for each attribute distinctly across the three class labels**

Cancer which is contrary to the general belief of a person being exposed to Lung Cancer only after a certain age.

Figure 4 is the distribution plots of independent features and hence the following facts can be inferred that for a significant number of plots, the grey density plot peaked when the value at the x-axis increases which means that when the value of the feature metric like lung disease, occupational hazard or chest pain increases as the level of risk of Lung Cancer also increases. However, it can be observed that in certain specific features that the peak of the density is not at a relatively high and one particular value. The features which do not follow the trend and are evident of a different trend are as follows:-

- For the case of an individual in the data set being a smoker, even if the habit of smoking is not at its worse, implying that the feature has a lower value on the X-Axis, it is not definite that the person does not have a high risk of Lung Cancer. A person is exposed to a medium risk of Lung Cancer even if the individual is not an active smoker.
- Even in the case of a passive smoker, it is evident that being exposed to smoking even at not a severe value increases the risk of Lung Cancer and is not definite for low risk of Lung Cancer.
- For attributes such as the measure of fatigue, severity of coughing blood, and rate of weight loss, it is not definite that a person is at high risk at a very high value of such parameters. Even at a low value of the discussed parameters, the person is exposed to a risk of Lung Cancer.
- For general symptoms such as the severity of shortness of breath, wheezing, swallowing, clubbing of fingernails, and frequent cold, there is not a definite peak for any of the classes and the prediction is not definite in the context of any of these parameters. However, the parameters support prediction for distinct classes of Lung Cancer.
- From the features which inhibit a definite peak, it can be easily inferred that people who consume less alcohol are at least at risk of having Lung Cancer and those who consume more alcohol have a high risk of Lung Cancer. Also, with an increasing level of the tendency of allergy to dust, the severity of the probability of developing Lung Cancer increases.
- It can be concluded strongly that males have a higher risk of lung cancer as compared to females.

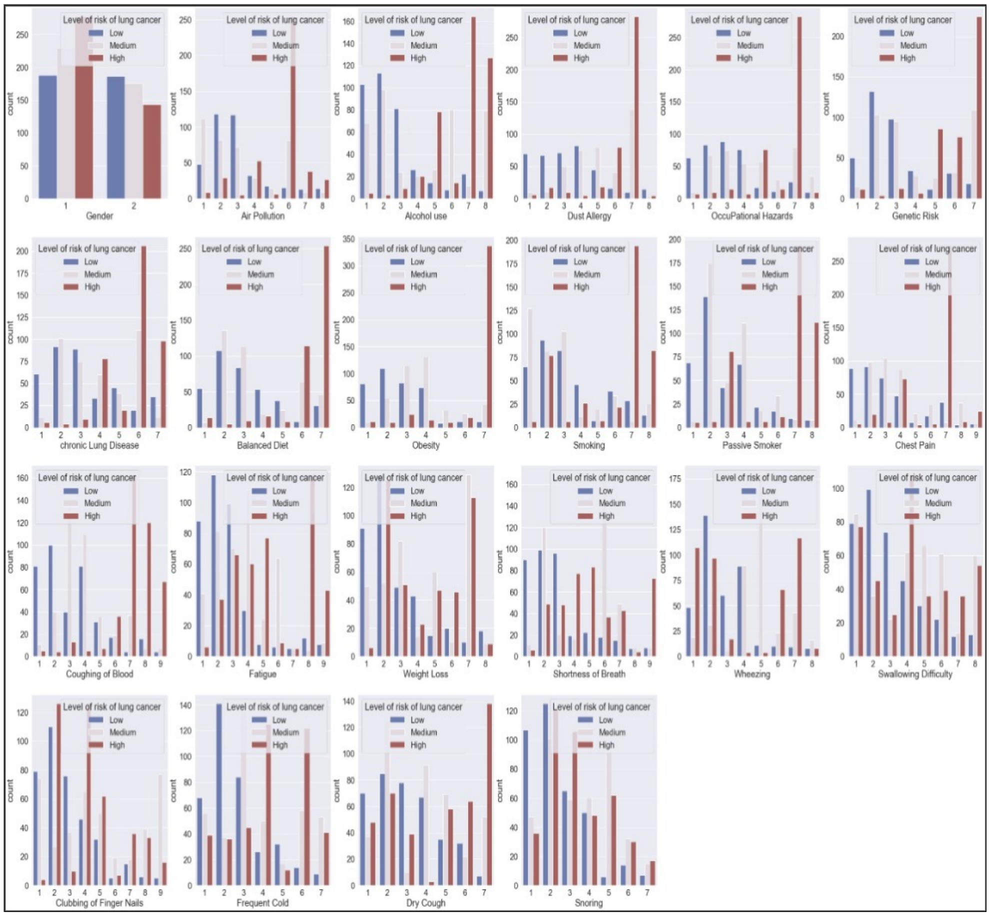


Figure 5: **Count-plot for each feature value to visualize if the attribute can clearly identify Low, Medium, or High Risk of Cancer at any specific attribute value**

From the count plots as in Figure 5, apart from inferring the levels of risk of lung cancer at various levels of the attributes or column values, it can be strongly observed that for the features which are not highly correlated with the target, there is no definite difference amongst the relative height of bar plots at any two distinct values on the X-Axis. Hence, only on the basis of these specific features in the dataset, it is not possible to strongly predict the level of risk of lung cancer is increasing or decreasing, whereas in contrast to such features, the attributes which are highly correlated with the target, support the process of strong prediction. For example, in passive smoking, if someone is a 7-level passive smoker, he/she has a high risk of lung cancer.

### 3.3 Feature Selection Techniques

Feature Selection is the process of recognizing and selecting a subset of input features that are most relevant to the target variable. The purpose of feature selection is to minimize redundancy in the dataset, cut the computational cost, and achieve an efficiently reduced dataset. Feature Selection can help find accurate data models. The proposed work demonstrates three different feature selection techniques.

#### 3.3.1 High Correlation

A correlation matrix is a table that displays correlations between all the input features. It is a powerful tool for summarizing a large dataset and identifying and visualizing patterns in the given data. Every cell of the matrix has a correlation coefficient. Heatmap can be used to visualize the correlation between variables. Using a heatmap, it becomes handy to read the correlation matrix.

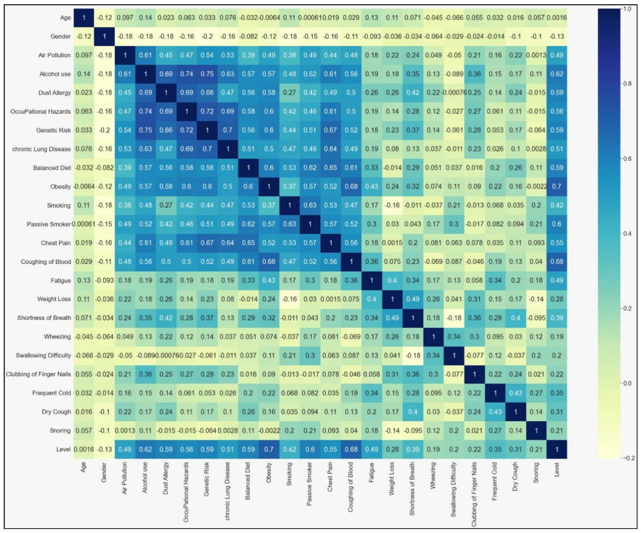


Figure 6: **Heat-map to visualize the correlation value amongst the features and the features and the target label class**

The closer the correlation coefficient is to 1, the more positively correlated features are. The closer the correlation coefficient is to -1, the more negatively correlated features are. For our dataset, the threshold for the correlation coefficient is 0.6. Suppose any input features have a correlation coefficient greater than 0.6 with the target variable/feature. In that case, it implies it is highly correlated with the target feature and can be considered for feature selection and model training.

### 3.3.2 Chi-Square Test

The chi-square [5] is used for feature selection to select the relevant features in the lung cancer dataset. It is commonly used in testing the connection between two categorical outcome attributes in feature selection. It allows us to test whether the attributes are independent or not. When two features are independent, we have a minimum chi-square value, and the outcome count approximates the expected value. The hypothesis of independence is false when the chi-square value is high. More chi-square value indicates that the feature is more dependent on response and can also be utilized for model training. After removing the feature with maximum p-value, we will train the model again to attain a model with good accuracy.

Chi-Square Formula

$$\chi^2 = \sum \left( \frac{(O_i - E_i)^2}{E_i} \right) \tag{1}$$

where $\sum$ = the sum of observed and expected value, O = observed Value, E = expected value

### 3.3.3 Sequential Feature Selection

Sequential feature selection is an iterative and greedy algorithm. It is a part of the wrapper method where it adds and removes features from the dataset sequentially. It evaluates each input feature and selects ’x’ features from ’n’ input features based on individual scores. Mathematically these algorithms are used to reduce initial ’n’ features to ’x’ features where x<n. and the ’x’ features are optimized for the model’s performance.

## 4 Classification Techniques

### 4.1 Logistic Regression

Logistic Regression [6] is one of the simplest machine learning algorithms, which is easy to implement, and interpret and very efficient to train. It is a classification model that employs

Table 1: **Feature Selection Techniques Analysis**

| Feature Selection Technique | No. of Features selected | Optimised Features | Best Model | Test Accuracy |
|---|---|---|---|---|
| High Correlation | 10 | Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Obesity, Passive Smoker, Chest Pain, Coughing of Blood, Gender | Random Forest | $88.42\%$ |
| Chi-Square Test | 10 | All features except “Age” | Decision Tree | $89.25\%$ |
| Sequential Feature Selection (SFS) | 12 | Air Pollution, Alcohol use, Dust Allergy, Obesity, Passive Smoker, Coughing of Blood, Fatigue, Weight Loss, Wheezing, Swallowing Difficulty, Frequent Cold, Snoring | Random Forest | $90.90\%$ |

the sigmoid function as a cost function in order to return a probability value that can be mapped to discrete classes. Logistic Regression parameters are estimated by maximizing the logarithmic likelihood function using training data.

Logistic regression can be run in these steps:

1. Calculate using the logistic function.
2. Learn the coefficients for a logistic regression model.
3. Finally, make predictions using a logistic regression model.

The Logistic Function is given as follows:

$$\frac{L}{1+e^{-K(x-x_0)}} \tag{2}$$

where e = Euler’s number, $x_0$ = Middle x-value of **Sigmoid Function**, L = The maximum value of curve, K = Abruptness of curve.

The logistic regression model is given in the equation as:

$$\frac{e^{b_0+b_1*x}}{1+e^{b_0+b_1*x}} \tag{3}$$

### 4.2 Support Vector Classification

SVM is one of the most popular classification algorithms with an elegant way of transforming nonlinear data [7]. It classifies the input dataset by introducing a hyperplane boundary that separates the dataset into two parts [8]. For non-linearly separable datasets, SVM is more suitable since it reduces the misclassification rate. SVM’s goal is to minimize an upper bound of the generalization error by maximizing the margin between the separating hyperplanes.

### 4.3 K-Nearest Neighbours

The KNN algorithm is a supervised classification method. It is a simple algorithm that looks for the nearest fit. The database is compared to the comparison set. The test sample’s mark is determined by the closest match of the k nearest neighbors. To calculate the distances between research samples and database samples, various distances such as Euclidean, cosine, and similarity are used. [9]

**Algorithm 1** Algorithm of K-Nearest Neighbours

```
Require: $X \leftarrow dataset$
$k \leftarrow number\ of\ neighbors$
$train, test \leftarrow split(X)$
while $True$ do
    calculate Euclidean distance of k neighbors
    $E_{distance}(k) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
    $Take\ nearest$ k $neighbors\ as\ per\ Euclidean\ distance$
    $Count\ no\ of\ data\ points\ in\ each\ category$
    $Assign\ new\ data\ points\ to\ that\ category\ for\ which\ no\ of\ neighbor\ is\ maximum$
end while
```

### 4.4 Random Forest

Random forests [10] or random decision forests are an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. The training algorithm for random forests applies the general technique of bootstrap aggregating or bagging to tree learners.

Given a training set X = $x_1, ..., x_n$ with responses Y = $y_1, ..., y_n$, repeatedly bagging (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

- Sample, with replacement, n training examples from X, Y call these $X_b, Y_b$.
- Train a classification or regression tree $f_b$ on $X_b, Y_b$.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' as shown below:

$$\frac{1}{B}\sum_{b=1}^{B} f_b(x^{'}) \tag{4}$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

### 4.5 Decision Tree

A decision tree [11] uses a supervised learning technique to build a model which is in the form of a tree data structure (set of nodes arranged in a hierarchical fashion). Initially, the entropy of the parent is calculated. The information gain is calculated by subtracting the weighted sum of the entropy of children from the entropy of parents. The one with the highest information gain is considered the root node and the process goes on until the classification is done. Given new test data, the tree is used to predict the result.

### 4.6 Naive Bayes

Naive Bayes [12] is mostly used in the area of Data Mining and Machine Learning. It is a statistical classifier that assumes no dependency between attributes but attempts to maximize the posterior probability in determining the class. Initially, in order to decide which class the instance belongs to, probabilistic value is calculated. The final class label is the class with the highest probability value. The principle of Naive Bayes is based on Bayes's rules of the simple conditional

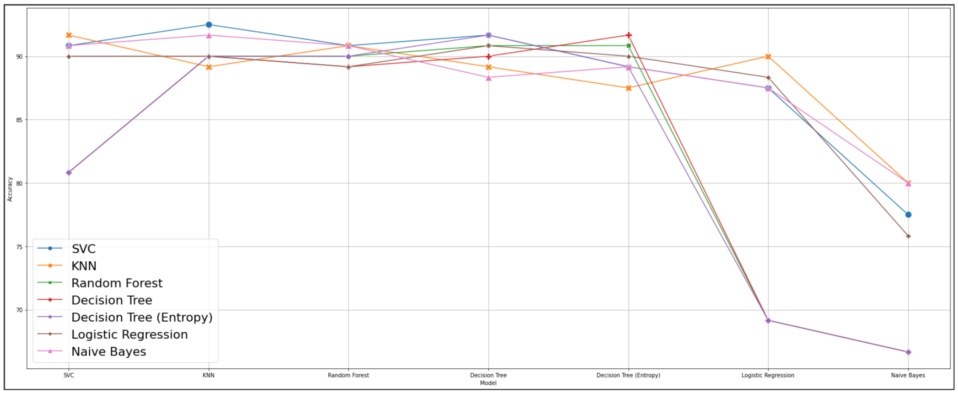


Figure 7: **The plot for visualizing accuracy for each Classification Technique used as an estimator. The best estimator, as visualized from the given graph for Sequential Feature Selection(SFS) is K-Nearest Neighbours(KNN)**

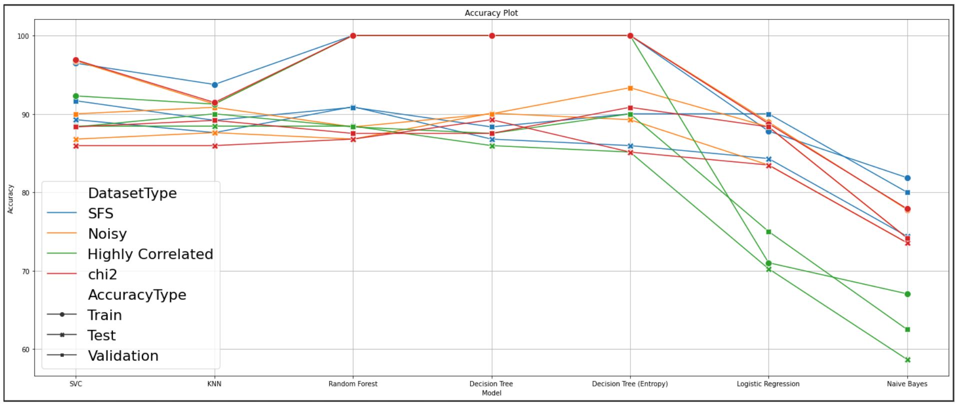


Figure 8: **The plot for comparison of the relative accuracy of predictions by each model on Train, Test, and Validation Data**

probability that is used to estimate the likelihood of a property given the small amount of training data to estimate parameters such as mean and variance necessary for classification.

## 5 Experimental Setup

### 5.1 Data Cleaning and Data Pre-processing

After loading the dataset, we checked for basic details like the number of rows and columns in the dataset, null values, and the count of each class label in the target column. From this, we know that the dataset contains no null values. Our target column contains three classes, i.e., low, medium, and high, which are the levels of cancer. The count of class low is **303**, the medium is **332**, and high is **365**. As only this column is of object/string type, we need to encode the target column into the numerical values, so we have encoded the target class label low - 0, medium - 1, and high - 2. We have checked that there is no outlier in the dataset. So for that, we have added some noise in our dataset of around 200 rows. After that, the total rows become 1200, and the columns remain the same as before. The class label count after adding noise becomes low **374**, medium **406**, and high **420**. The ***min-max normalization*** method is used to normalize the data. The data are re-scaled using the min-max method between 0 and 1. The machine learning algorithm is trained more effectively because of the uniform data. For the same reason, there is a need to normalize our data. The data is split into the ratio of Train, Validation, and Test as 80:10:10.

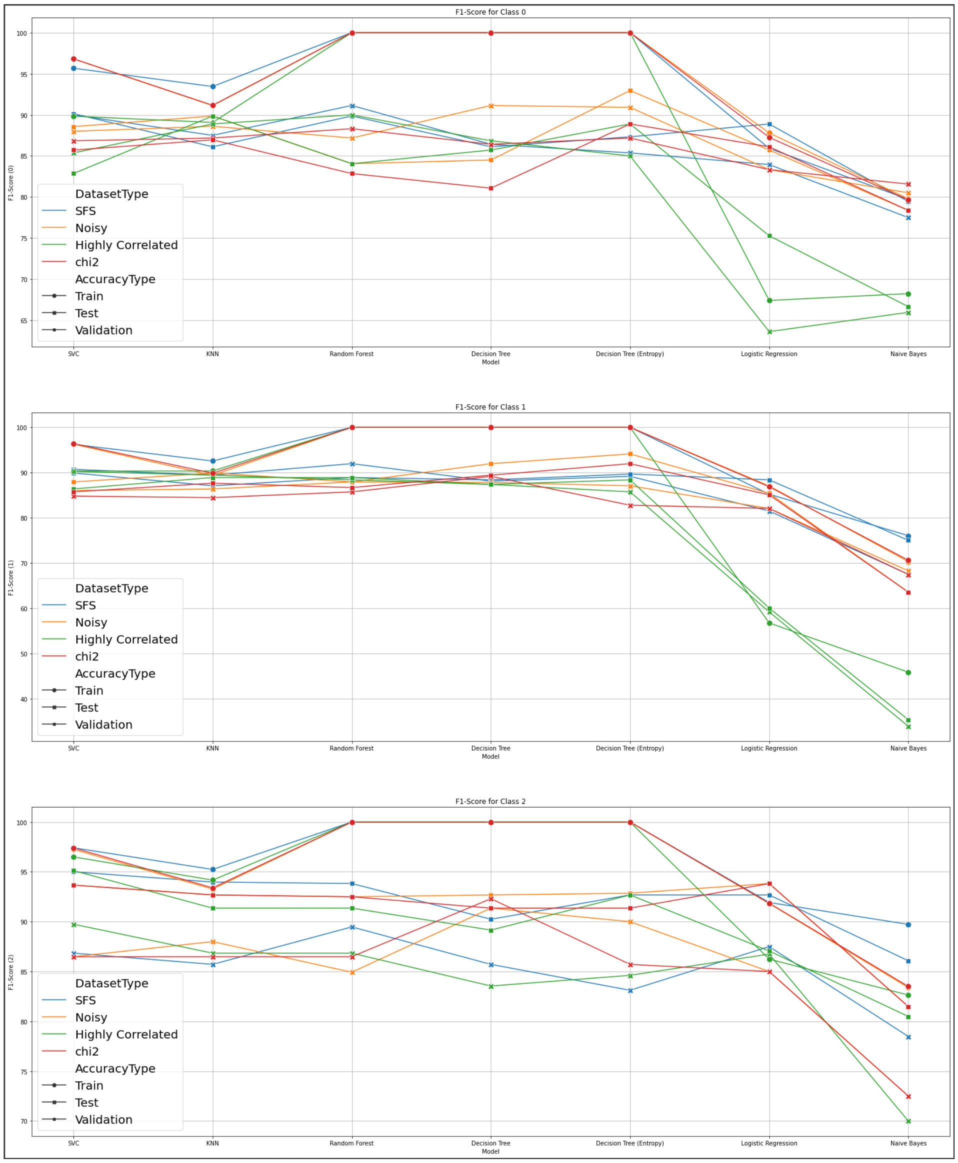


Figure 9: **Plot for relative F-1 Scores for Train, Test, and Validation Datasets for each Class**

### 5.2 Hardware and Software Specifications

The project is designed and developed using Python and its libraries as Numpy, Pandas, Sci-kit, Matplotlib, Seaborn and Plotly. Google Colaboratory is used to develop and run the code on the Web Browser itself. The specification for the platform are as for GPU is 1xTesla K80, having 2496 CUDA cores with 12GB GDDR5 VRAM. The CPU specifications are as 1xsingle Core Hyper-Threaded Xeon Processors @2.3Ghz i.e(1 core, 2 threads) and the available RAM is approximate 12.6 GB with a disk space of almost 33 GB.

## 6 Results and Discussions

This section briefly discusses the evaluation metrics and the value of the metrics retrieved from implementing different classification techniques.

### 6.1 Evaluation Metrics

The evaluation metrics used to ascertain the performance of each classification model are Accuracy and F-1 Scores. The value for each evaluation metric is calculated distinctly for Train, Validation, and Test Data. With an aim to add definiteness to the final result, the class-wise F-1 Score is calculated for each of the classes, as shown in Figure 9. The model which performs consistently across different feature selection techniques while optimizing on time and ensuring its suitability for large amounts of data, is considered as the best for the task. Additionally, the goal is to verify that the performance of easily explainable models based on these metrics is improved as the model is supplied with a subset of features. The subset of features is curated based on feature selection techniques, as discussed above.

### 6.2 Results

The highest values for test accuracy reported for the task of prediction of risk of Lung Cancer are presented in this section for each of the classification techniques along with the process of feature selection, which led to the result. For the case of the Decision Tree Classifier, the results reported on Noisy Data are as 90.08%, 90.90%, 87.80%, and 91.56% for accuracy, F-1 Score for Class 1, F-1 Score for Class 2, and F-1 Score for Class 3 respectively. The values for the Decision Tree Classifier with Entropy as the criterion of split and the feature selection technique as Sequential Feature Selection (SFS) are 88.42%, 86.74%, 92.5%, and 86.07%, respectively. For Random Forest Classifier, implemented as an Ensemble Learning Technique, the values are 89.25%, 91.13%, 89.41%, and 87.17% when the model is trained and tested with a subset of features selected based on high correlation. For the case of a Support Vector Classifier, the corresponding values are 89.25%, 90%, 90.69%, and 86.84% by using SFS, and the same technique delivers appreciable results for classifiers such as Logistic Regression and Naive Bayes, the values for evaluation metric reported are as 84.29%, 83.95%, 81.48%, 87.5% and 74.38%, 77.5%, 67.46%, and 78.48% respectively. The corresponding values for K-Nearest Neighbours(KNN) with highly correlated features are 88.42%, 88.88%, 89.41%, and 86.84%.

## 7 Conclusion

It can be inferred from the results that Sequential Feature Selection (SFS) technique, improves the result over test data for multiple classification techniques. It can be strongly concluded from Figure 7, for the case of Sequential Feature Selection(SFS), the best estimator is KNN as the subset of features selected by this estimator improves the accuracy for relatively easier classification techniques such as Logistic Regression and Naive Bayes. For the same reason, as can be observed from Figure 8, the test accuracy for Logistic Regression and Naive Bayes is approximately 84.29% and 74.38% respectively which is a significant improvement over the values of 70.24% and 58.67% for when the models are supplied with highly correlated features on the basis of the correlation matrix. Hence, even though the Decision Tree Classifier performs well on the data to reduce the time complexity and enhance the explain-ability of the techniques used for classification, the Sequential Feature Selection technique can be used for effective results.

## References

[1] Jaiman, Hemant, Kuldeep Sharma, and K. Sujatha. "Survey on lung cancer detection using machine learning." International Journal for Research in Applied Science and Engineering Technology 8.6 (2020): 1970-1973

[2] Dharmarajan, Adarsh and T. Velmurugan. “Lung Cancer Data Analysis by k-Means and Farthest First Clustering Algorithms.” Indian journal of science and technology 8 (2015).

[3] Zehra Karhan1, Taner Tun,” Lung Cancer Detection and Classification with Classification Algorithms” IOSR Journal of Computer Engineering 18(6), Ver.3 Dec 2016, PP 71-77.

[4] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138.

[5] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 160-163, doi: 10.1109/ICSESS.2018.8663882.

[6] R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.

[7] B R Manju et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1012 012034

[8] S. S. Raoof, M. A. Jabbar and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 108-115, doi: 10.1109/ICIMIA48430.2020.9074947.

[9] Mustafa Abdullah, D., Mohsin Abdulazeez, A., & Bibo Sallow, A. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. Qubahan Academic Journal, 1(2), 141–149. https://doi.org/10.48161/qaj.v1n2a58

[10] N. Banerjee and S. Das, "Prediction Lung Cancer– In Machine Learning Perspective," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132913.

[11] R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.

[12] R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.