# Lung Cancer Risk Prediction Features Influence Model based on Machine Learning Techniques

Subhash Mondal
*Computer Science and Engineering*
*Central Institute of Technology Kokrajhar*
Kokrajhar, India
ph22cse1001@cit.ac.in

Ranjan Maity
*Computer Science and Engineering*
*Central Institute of Technology Kokrajhar*
Kokrajhar, India
r.maity@cit.ac.in

Chirag Rai
*Computer Science and Bussiness System*
*Meghnad Saha Institute of Technology*
Kolkata, India
chiragraicr7@gmail.com

Souptik Pramanik
*Computer Science and Bussiness System*
*Meghnad Saha Institute of Technology*
Kolkata, India
souptikpramanik9933@gmail.com

Amitava Nag
*Computer Science and Engineering*
*Central Institute of Technology Kokrajhar*
Kokrajhar, India
amitava.nag@cit.ac.in

***Abstract*—Currently, lung cancer is a very common form of cancer. This is because many people are chain smokers nowadays, and many are affected due to their work hazards. The pollution level in modern cities is also a major cause of this type of cancer. This model is built to predict the chances of the occurrence of lung cancer in an individual with the help of certain conditions. The acquired dataset used in this study contains multiple features, but not all are necessary for predicting the risk of lung cancer. Hence, an embedding feature importance model Light Gradient Boosting Machine (LGBM) is used to find the impact of every feature, and the model had trained using the features with maximum influence. The dataset has been divided into two parts for training and testing the model. The models achieve a k-fold mean accuracy of 97.63% and above with all the features and more than 93% on the reduced features for all the deployed models. The models are developed based on a resource-constrained device perspective over the reduced features low resource dataset and use an algorithm to measure the execution time taken for every model to run and complete its prediction after fitting with the respective classifiers. The model developed on medical data should have maximum accuracy and is necessary for time efficiency that reflects on all the deployed models with stability and efficacy, indicating the robustness and non-overfitted model.**

***Keywords—Lung Cancer, Light Gradient Boosting Machine (LGBM), Feature Influence, Prediction Model, Machine Learning***

## I. Introduction

Lung cancer is currently observed in masses and comes after skin cancer in terms of common cancers worldwide. A study in 2020 found that there are over 2 million new patients with severe lung cancer [1], and the average age of occurrence is 55-65. Hungary, Serbia, and France (New Caledonia) are the top three countries with the highest rate of new lung cancer cases, according to the same study in 2020. The study found that men are the common victims of this cancer hence it is the most common type of cancer in men. In the case of women, lung cancer comes after skin and breast cancer. There are several reasons why lung cancer is developed in an individual. The most common reason is active and passive smoking, followed by inhalation of polluted air (poor AQI), a person's genetics, occupational hazards, dust, smoke from factories, or burning waste [2]. Lung cancer can be avoided with a healthy lifestyle by consuming healthy and vitamin-rich meals and avoiding smoking (both active and passive). So, the model is built to predict the chances of occurrence of lung cancer in an individual with the help of certain conditions like smoking, both active and passive, etc.

The model approaches the problem uniquely using feature selection, as the number of columns is too many. Hence, we need to manually select a few features which are necessary for the model to predict a sample input with the highest accuracy and precision. For this, we need to use features that suggest a better result that can be computed using LGBM Classifier. Heatmap is used for calculating the weightage of each feature against the other. The model also has performance matrices such as k-fold Mean Accuracy (Ma), Validation Accuracy (Acc), Recall (Re), Pression (Pre), and F1-Score (Fs), Cohen Kappa Score (Cks) using the confusion matrix and classification matrix. The standard deviation (Sd) is used to check the data correlation and bias of the model, and the Execution Time (ET) measures the model responsiveness. For developing the model, we use machine learning (ML) models such as boosting classifiers like Hist Gradient Boost (HGB), LGBM, XGBoost (XGB), CatBoost (CB), Gradient Boost (GB), and few popular algorithms like K-Nearest Neighbor's (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). Execution time is also measured for every model with graphs of different values. Graph of actual values v/s predicted values of every algorithm is plotted to understand the model's fitting.

In this study, we have taken a multiclass dataset consisting of 1000 data items. The risk of lung cancer occurrence can be of three types, namely high, medium, and low, based on certain conditions of an individual. This study is focused on the chances of occurrence of lung cancer in an individual and not the prediction model of lung cancer detection. Boosting algorithms achieve very high accuracy values. In most models, 100% accuracy is observed except for the LR and SVM classifiers. DT, KNN, LR, and SVM complete their executions before feature selection in less than a second. After using the features selection model, we achieve similar results with XGB and CB Classifier, pulling off less than a second in execution.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

The study is in sections; Section II illustrates the related works, followed by the proposed methodology in Section III, and concludes with the future scope discussed in Section IV.

## II. Related Work

This section considered the related articles on lung cancer detection and prediction model that have already been published, but not many articles are found to concentrate on the risk of lung cancer prediction feature influence model that concludes a decision regarding which features are contributing more to the risk of this disease. Some related literature is discussed in detail, along with their claimed experiment results on a particular deployed model or a combination of models.

In [3], a lung cancer malignancy classification model is built mainly with various ML algorithms. The difference between benign and malignant tumours was clear in this study. Mostly various ML algorithms like KNN, DT, SVM, and NB are used; KNN with 72.24%, SVM with 62.5%, NB with 62.5%, and DT with 58.11% accuracy was achieved. Apart from these, Sd, Pre, ROC, AUC, and Fs are measured for each ML model. Extensive research is done to reduce death due to lung cancer. Research has been done carefully because a small error may lead to an erroneous result. In [4], an ML model is built to detect lung cancer with the help of image processing. Several ML techniques like SVM, KNN, clustering method, NB, and deep learning (DL) techniques like Convolution neural networks (CNN), Deep Neural Network (NN) methods were used. The SVM-KNN method shows 95.56%, Improved NN 94.58%, CNN 94.56%, unsupervised learning like clustering shows 93.91%, and SVM and NB achieved 98.3% accuracy. DNN is a strong tool for detecting lung cancer which was used in this model. In [5], the paper suggests a study about the different ML techniques used while working with a dataset based on Lung Cancer. The study shows the multiple papers that use different algorithms and predicts the values based on the model's training on different datasets. The author uses other research papers to understand and compare their working module with all types of pre-processing methods, classification methods, training, and testing models. The paper presents the quantities of data used for every other model developed by the authors. The author talks about the benefits of data cleansing, normalization with feature selection, and extraction using different ML models. The document shows the maximum accuracy of every model in each paper and presents the data split for training and testing of all the papers in the conversation. Finally, the author discusses the future development of a prediction model for lung cancer with high efficiency and accuracy. In [6], a supervised learning model is built using various ML techniques to identify lung cancer in an individual with the help of certain symptoms. In recent time lung cancer has been very common in middle-aged people. Smoking and living in a polluted area are the most common factors for lung cancer and other lung diseases. Several ML algorithms are used, like SGD, Bayes Net, NB, SVM, LR, KNN, RF, Random Tree, AdaBoost (AB), and DL techniques like ANN. After that, an evaluation like Re, Fs, Pre, Acc, and AUC shows 99.3%, and the F1 measure shows 97.1% accuracy. This study can be used in the future and will be upgraded; LSTM and CNN can be done. In [7], the paper suggests a model that compares lung cancer's effect on men and women by using image-based analysis for predicting Lung cancer. This imagery uses lung data from CT scans, MRI, PET, and X-Ray medical images. The images are used for extracting data such as statistical and textual, but the extracted data is multiclass and imbalanced hence, SMOTE is used. This makes all 5 classes at the same level of 270 data samples. ML techniques like DT, RF, KNN, and AB are used as all the classes are balanced. The results of most of the models are 87% of accuracy and 73% of F1-Score. The ROC-AUC curve gives a better graphical representation of the model's values.

In [8], the author uses a low-resource database with only 309 rows and 14 columns, resulting in binary values found on the Kaggle website. Then the dataset is divided into a 0.67:0.33 ratio for training and testing using the train test split technique. Microsoft's Power BI is used for updating the database regularly so that this project can also have applications in real life this is also possible as it is defined on a small database, so it does not require many resources. ML models used for this project are LR, DT, RF, and KNN. The results are impressive as they achieve more than 85% accuracy for each model, whereas RF touches the highest 96% accuracy. In [9], the paper investigates the detection of lung cancer among women and men using a database collected from the surveillance, epidemiology, and end results database. Hence the study is done using multiple very popular ML algorithms to find the survival rate of women and the chances of detection of lung cancer in different genders using the data of 28000 plus patients. They have used multiple classes and columns to do pre-processing and build a binary classification algorithm using ML such as NB, DT, RF, XGB, KNN, LR, and SVM. For data analysis techniques such as Chi-squared and T-test have been used for measuring relationships between features and gender of the patients. Cox regression is used to univariate the analysis of survival in each feature. The results show more chances of women surviving than men and the results show that the male for single year would survive only with 85.9% whereas the women will survive with 92.4%. Similar difference in results is seen while comparing for 3 to 5 years of lung cancer. The ML models also performed well with the XGB achieving the highest accuracy of 90.75% for single year problem and LR achieved 75.65% for 3 years and 71.19% for 5 years. In [10], the author proposes a theory which states that predicting cancer could be easier if we can understand whether the tumour inside the lungs of a patient has mutated itself into cancer or not. This prediction can save a lot of time for the doctor which is essential in the diagnosis of such disease. For this we can use the images corresponding to the patients and extract certain features from it. This can be used to for a table with multiple columns for that we would need clear images which is not possible as medical documents have noise in them and we need to clear the noise using Digital Image Processing. Further, the features collected from the processed images will be used for developing a model with the help of ML classification algorithms such as SVM, RF, and ANN. The accuracy for the ML algorithms which work on the features extracted from the processed images is more than 80% and a Pre of 100% with Fs of more than 80%. In [11], the author talks about increasing the accuracy of early detection of lung cancer with the help of automation and eliminating human involvement. It is a comparative study done using ML algorithms such as RF, SVM, LR, and DT. The task has been divided into 4 parts the first is a method of combining treatment using radio waves with a KNN classifier, and the second is a tailored network of CNN, namely AlexNet, Resnet101, Inceptionv3, and Inception Resnet v2. The third algorithm is a fusion of CNN with the LSTM network, and the last model is a combination of LSTM, CNN, and radionics. All four ML models show more than 88% accuracy and more than 90% precision.

In [12], the paper canvases using Lung scan analysis reports for techniques such as X-Ray, CT scan, or MRI for categorizing the type of lung cancer a patient is suffering from using different image processing techniques. This can be done with the help of multiple ML algorithms, mainly classifiers such as KNN, SVM, DT, NB, Gradient Descent, MLP (Multi-Layer Perceptron), and RF. The biggest challenge for the author is the dataset's huge size, containing 15,000 images. The resources required for the task are huge hence it is divided into two parts image processing and classification. The accuracy of the KNN, RF, MLP, and DT algorithms is more than 80%, and Fs is more than 0.80. In [13], the paper discusses the application of AI in the field of healthcare to frame a computer-supported diagnosis system with the core of ML models. The dataset has 16 columns and 309 rows and has been acquired from Kaggle. The author has designed the mechanism using seven models: the IBL (Instance-Based Learner) Classifier, AB, J48 classifier, Logit Boost Classifier, LR, SVM, and RF classifier for early prediction of Lung Cancer using research paper. Machine Learning models are used with parameters for feature selection with Principal Component Analysis and Gini Index Technique. All the models achieve accuracy from 90% to 93%.

Above stated literature shows that there is more scope to explore the risk prediction of lung cancer using the ML model and also find out the effect of features that contributed more to the risk of this disease. Feature influence determination by the embedded feature selection ML algorithms is a new scope and approach for the risk prediction model.

## III. Proposed Methodology

This segment discusses the details of the work process executed throughout the dataset. Details about the dataset, like its acquisition, graph plotting, and other related information, have been illustrated in this portion. Techniques used to pre-process the dataset and various ML models' implementation are also reviewed. The environment used for this analysis is Python 3.8.

### A. *Dataset Acquisition & Description*

The dataset accommodates data on patients' chances of lung carcinoma, which is influenced by a few factors such as their age, gender, chest pain, air pollution exposure, alcohol use, dust allergy, and some others, as given in Table II. The table also gives detailed information about the dataset and its features. The dataset contains 26 columns. The 2 columns were dropped due to non-significance, of which 23 features have been given a numerical value, and the target column 'Level' is multiclass based on a study done on 1000 patients for over 3 years.

### B. *Data Pre-processing*

For data to be processed accurately, there should be a measure to calculate the precision of the model's prediction, and for this, we need multiple independent variables, which, when calculated together, give value to a dependent variable. This process can be done very easily using feature engineering. This helps in amending features and customizing the features important to the model. This feature can also be used to select custom features depending on the manual requirements of the user. The model can also use these custom-selected features for training and prediction. The dataset has all values in numerical form except for the dependent variable, which is in text and needs to convert to numerical value for analysis. Label encoding is the technique used to translate categorical data into numeric values to be machine-readable. The column of 'Level' in the dataset is the only column with a categorical value. Therefore, this column must be converted to numerical form, which is done using the label encoding technique. The feature correlation heatmap among the features and with the target column is represented in Fig. 1.

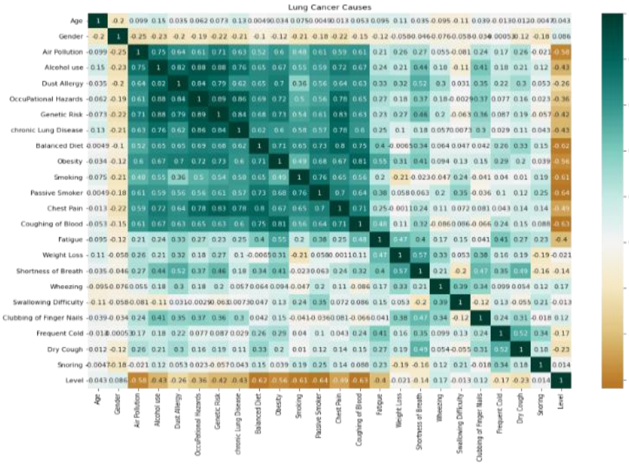

Fig. 1. Correlation Heatmap of the dataset

The dataset contains more than 1000 rows and 26 columns which specifically focus on the report of a person and is related to their lifestyle. This dataset has been collected for patients considering their daily interactions and way of living. But the dataset does not have any null values in it. This information is gained by using the info () function, which tells about all the data values in the dataset and their data type and states that there are no null values in the dataset.

The essential requirement for developing any model in supervised learning is to have a numerical value for making the model understand the dataset. Although the data contains mostly numerical values for all the columns, the most important target column is in string form, complicating the problem. To solve this complexity, we use an algorithm technique of pre-processing modules from the sklearn package. This algorithm is known as Label Encoder.

The dataset has three classes that have been used to define the magnitude of the probability of a person being diagnosed with Lung Cancer. Although the dataset has a small number of patients in consideration given here, the classes are very well balanced with no bias in patients of any one class. The information is gained by using the value counts () function, which shows that the classes 'High' with 365 values, 'Medium' with 332 values, and 'Low' contains 303 patients. But we do use Oversampling technique for balancing the dataset with SMOTE (Synthetic Minority Oversampling Technique), which increases the data values for each class in a stable way. The class with the maximum value is selected as a target to acquire, which is done using SMOTE and is achieved in the instances for both the other classes.

The dataset needs to be modified to the scale with features necessary for model training, making normalizing paramount. The data values vary in a difficult range to put on a scale, so we use the *StandardScaler* technique. The *StandardScaler* technique scales the data points by deducting the average calculated from all the points in the column. It tries to keep the values close to 1 or 0 with positive or negative signs.

### *C. Dataset division*

The dataset, after processing, is divided into two parts of 0.80:0.20 ratios, with 80% going in favour of training the model while the 20% testing the model developed using the ML classifiers. This split is necessary for the dataset as it will only assure the accuracy of the dataset and would also help in the testing of the dataset without searching for new values. This data can be used not only for measuring the accuracy but also for measuring the k-fold accuracy, which gives a mean value of accuracy and is used for measuring accuracy throughout the dataset.

### *D. Model Training & Testing*

The independent variable in the dataset is the features, and the dependent variable is the correct output of the given dataset, the ‘Level’ column which consists of three classes high, medium, and low. These three classes are the probabilities of people having different chances of suffering from lung cancer. These values can help develop a model using multiple ML models for predicting a person's chance of developing lung cancer. We have used multiple ML models for processing and analyzing the dataset, and these algorithms are used for creating a model from the dataset. This is done using ML models on the dataset, such as Classifiers like RF, DT, KNN, and SVM, and boosting algorithms like XGB, GB, HGB, CB, and LGBM. LR model is also used to find the dependent variable's accurate and precise numerical value. There is no requirement for parameter tuning for the evaluation as the model has high accuracy at default parameters, and further tuning might make the models overfit and increases the execution time with complexity. The results for all the deployed models are given in Table I.

TABLE I. EXPERIMENTAL RESULTS OF THE DEPLOYED MODEL

| Model | Acc | Ma | Sd | Pre | Re | Fs | Cks | ET |
|---|---|---|---|---|---|---|---|---|
| DT | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.19 |
| RF | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 |
| XGB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.10 |
| GB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 6.20 |
| HGB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 25.03 |
| LGBM | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.38 |
| CB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 |
| KNN | 100 | 99.75 | 0.0031 | 1.00 | 1.00 | 1.00 | 1.00 | 0.46 |
| LR | 97.50 | 99.63 | 0.0050 | 0.98 | 0.98 | 0.98 | 0.96 | 0.51 |
| SVM | 96.50 | 97.63 | 0.0108 | 0.97 | 0.97 | 0.97 | 0.95 | 0.67 |

### *E. Feature Important Model Training & Validation*

Furthermore, we evaluated the dataset after manually selecting the best-required features for the model prediction. For this, we used the LGBM Classifier, which helps us evaluate the importance of each independent variable on the dependent variable in numerical format.

The LGBM classifier is a boosting algorithm based on an inclination framework that uses a decision tree to surge the efficiency and accuracy of the model. The model is low memory based and uses two novel techniques i.e., Gradient-based One Side Sampling and Exclusive Feature Bundling.

The model uses all the columns given in the dataset with numerical values for training and prediction. But for designing a model with pinpoint accuracy and precision, we need to filter out features that are utterly necessary for our prediction model. As many columns in the dataset do not have enough importance and impact while predicting the results. So, we use an LGBM Classifier, which uses an algorithm to measure the impact of every feature in the dataset. Hence, we can use only those features and prevent any errors while making predictions and improve accuracy by a huge margin. The results in Table II present the numerical values for the impact of every column on the deployed model’s prediction algorithm. This data is also present in graphical format in Fig. 2. The features with a certain value can be filtered out to increase efficiency, hugely improving our results. For feature selection, we use a specialized parameter, Gini impurity, a concept used as a default for the development of good decision trees based on the dataset, and this is done using a simple mathematical formula.

TABLE II. THE GINI IMPURITY VALUES OF EACH FEATURE

| # | Features | Effect | # | Features | Effect |
|---|---|---|---|---|---|
| 1 | Age | 2581 | 13 | Wheezing | 245 |
| 2 | Air Pollution | 1026 | 14 | Swallowing Difficulty | 240 |
| 3 | Snoring | 540 | 15 | Chronic Lung Disease | 240 |
| 4 | Obesity | 468 | 16 | Weight Loss | 179 |
| 5 | Fatigue | 465 | 17 | Clubbing of Finger Nails | 157 |
| 6 | Dust Allergy | 455 | 18 | Balanced Diet | 134 |
| 7 | Passive Smoking | 342 | 19 | Gender | 132 |
| 8 | Coughing of Blood | 286 | 20 | Occupational Hazards | 102 |
| 9 | Genetic Risk | 276 | 21 | Shortness of Breath | 72 |
| 10 | Frequent Cold | 269 | 22 | Dry Cough | 62 |
| 11 | Alcohol Use | 263 | 23 | Chest Pain | 48 |
| 12 | Smoking | 261 | | | |

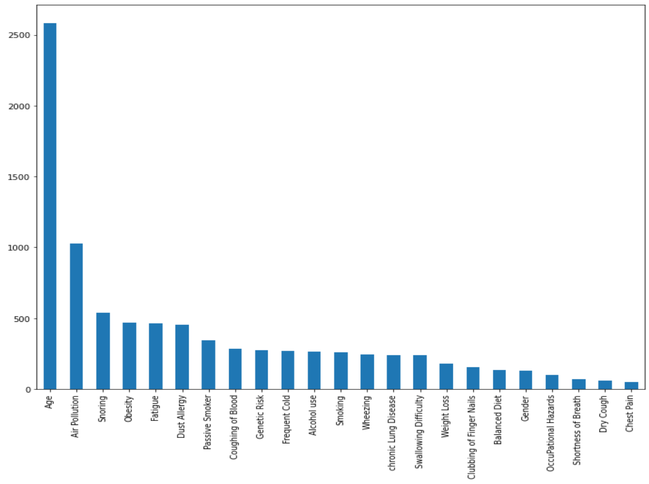

Fig. 2. Feature importance graphical representation

The set of features used for the analysis and development of a new model based on certain features which have been manually selected is given in Table III.

TABLE III. THE SELECTED FEATURE IMPORTANCE DESCRIPTION

| # | Feature's name | Description |
|---|---|---|
| 1 | Age | Demography shows that people between 55 to 84 have a maximum percentage of Lung cancer cases [14]. |
| 2 | Air Pollution | Pollution particles affect every one in ten persons suffering from lung cancer [15]. |
| 3 | Alcohol Use | Alcohol use comes in combination with other carcinogenic products [16]. |
| 4 | Dust Allergy | Dust blocks the air filter and prevents airflow in the body, which increases the growth of the tumour in the lung [17]. |
| 5 | Genetic Risk | A person's genes increase the chances of cancer diagnosis in an individual as they are mutated [18]. |
| 6 | Obesity | High-fat level in the body increases cell production, and this causes growth in tumour cells, causing lung cancer [19]. |
| 7 | Smoking | It is the primary cause of lung cancer as it damages the lung with smoke and tobacco [20]. |
| 8 | Passive Smoker | Smokers exhale harmful toxic chemicals like the smoke they inhale. This smoke, if inhaled, can impact the lungs and cause lung cancer [21]. |
| 9 | Coughing of Blood | Blood in a cough (spit) indicates a person has been diagnosed with lung cancer [22]. |
| 10 | Fatigue | Almost 80% of cancer patients exhibit fatigue sometimes due to chemotherapy [23]. |
| 11 | Frequent Cold | Frequent cold with a combination of different symptoms also potentially signifies [24]. |
| 12 | Snoring | Snoring affects the sleep apnea of a person, which can affect the person's health [25]. |

The experimental results of all the performance metrics of every ML model used for developing the new model with manual features is given in Table IV.

TABLE IV. THE EXPERIMENTAL RESULTS OF THE FEATURE IMPORTANCE MODEL

| Model | Acc | Ma | Sd | Pre | Re | Fs | Cks | ET |
|---|---|---|---|---|---|---|---|---|
| DT | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.17 |
| RF | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.24 |
| XGB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.84 |
| GB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 2.84 |
| HGB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 3.61 |
| LGBM | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.07 |
| CB | 100 | 100 | 0.0000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 |
| KNN | 99.50 | 99.75 | 0.0031 | 1.00 | 1.00 | 1.00 | 0.99 | 0.20 |
| LR | 96.00 | 96.63 | 0.0211 | 0.96 | 0.96 | 0.96 | 0.94 | 0.51 |
| SVM | 92.50 | 93.00 | 0.0165 | 0.92 | 0.93 | 0.92 | 0.89 | 0.67 |

### *F. Experimental Results Discussion*

The model's results trained on all the features given in the dataset achieves 100% accuracy for only a few models, such as HGB, LGBM, RF, and GB Classifier. Other models also have remarkable accuracy. In the field of medical diagnosis, it is very necessary to achieve maximum precision, the highest specificity, and clear-cut results. The lower the standard deviation, the better the model's efficiency, which also helps predict the result closest to accuracy.

The graph for every model, which plots the values of predicted and actual values also signifies the efficient working of the model. This also shows that the model is not overfit or underfit. It can also be used for understanding the working of a model and this makes it a great performance matric. The time function also works as a performance metric and is used to check the speed of prediction of every model and the fitting of model from the starting and until other performance scores are measured. The confusion matrix and the predicted curve of the LGBM model is depicted in Fig. 3.

We achieve our objectives by using the LGBM classifier that specializes for measuring the impact of every feature in our model. These results are used for selecting features which have the highest impact and this helps some of our models achieve maximum efficiency. After evaluating the final scores for both the models, the manual tuned model is working more accurately and predicting values more precisely and efficiently as it requires a smaller number of features and has an accuracy very close to 100%. Not only the accuracy score but also other scores such as k-fold mean accuracy, recall, cohen-kappa score, and F1- scores are remarkable.

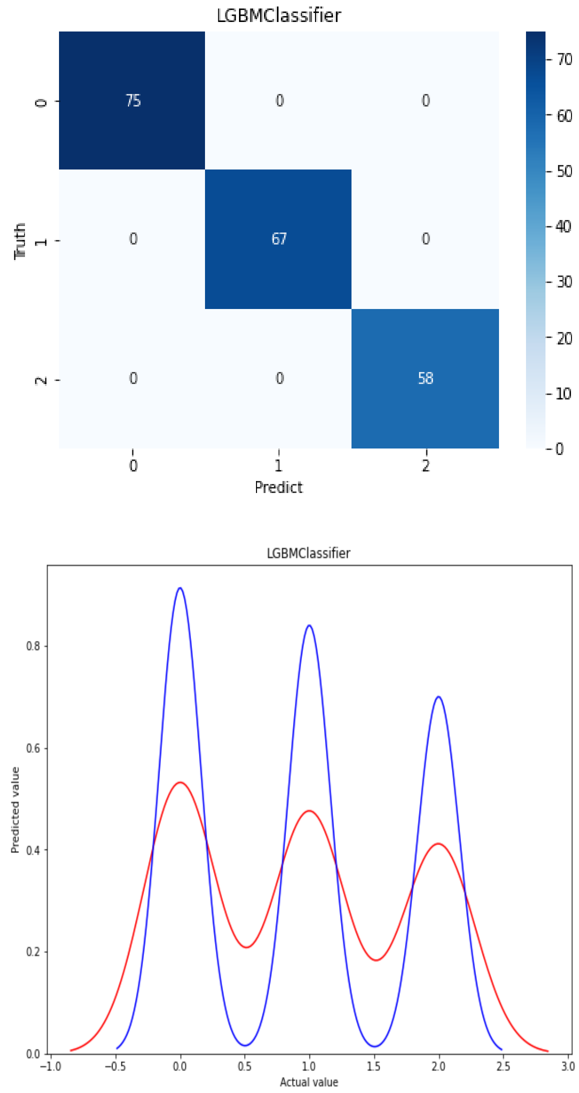

Fig. 3. The LGBM confusion matrix and prediction curve

## IV. Conclusion & Future Scope

This study attempts to build a low-resource model based on a dataset of multiple features but only a thousand rows. This is necessary for doctors who use devices with smaller processing power and is also very efficient for predictions. This model can also be deployed for mobile computing devices, which are abundant worldwide. The model will only improve by increasing the data inputs. But the model will only increase the efficiency for every new data input. The model uses a specialized algorithm that can be run by selecting very few features that are necessary for the model, which is already low resource, and this can make the model more efficient and accurate. Furthermore, the features the embedded LGBM model selected are the biggest indicator of lung cancer patients. This is also verified using the references given for every feature selected, making it medically valid. The experimental results create a new milestone for the risk of lung cancer prediction on the reduced feature model using LGBM or any other deployed ML classifiers. The execution time of the deployed models is very less from the perspective of the resource-constrained device in case of this low resource medical dataset; the lower standard deviation and high cohen kappa score indicate non overfitted and stability of each model. This study concludes by incorporating the mean accuracy, which is very effective and high in the multiclass based lung cancer risk prediction model. The study will be further extended with the addition of deep learning approaches.

## References

[1] W. H. O. "Lung cancer statistics," 23 March 2022. [Online]. Available: https://www.wcrf.org/cancer-trends/lung-cancer-statistics/.

[2] M. Markman, "Lung cancer symptoms," 12 September 2022. [Online]. Available:https://www.cancercenter.com/cancer-types/lung cancer/symptoms. [Accessed 25 January 2023].

[3] E. S. Neal Joshua, M. Chakkravarthy and D. Bhattacharyya , "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study," *Revue d'Intelligence Artificielle,* vol. 34, no. 3, pp. 351-359, 2020.

[4] G. Paliwal and U. Kurmi, "A Comprehensive Analysis of Identifying Lung Cancer via Different Machine Learning Approach," in *10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, MORADABAD, India, 2021.

[5] V. N. Jenipher and S. Radhika, "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques," in *3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020.

[6] E. Dritsas and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," *Big Data Cogn. Comput,* vol. 6, no. 4, p. 139, 2022.

[7] K. Ingle, U. Chaskar and S. Rathod, "Lung Cancer Types Prediction Using Machine Learning Approach," in *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2021.

[8] A. E. Celik, J. Rasheed and A. Yahyaoui, "Machine Learning Approaches for Lung Cancer Prediction," in *12th International Conference on Advanced Computer Information Technologies (ACIT)*, Ruzomberok, Slovakia, 2022.

[9] Y. Wang, S. Liu, Z. Wang and Y. Fan, "A Machine Learning-Based Investigation of Gender-Specific Prognosis of Lung Cancers," *Medicina (Kaunas),* vol. 2, no. 99, p. 57, 2021.

[10] N. Banerjee and S. Das, "Prediction Lung Cancer– In Machine Learning Perspective," in *International Conference on Computer Science, Engineering and Applications (ICCSEA)*, Gunupur, India, 2020.

[11] S. Agarwal, S. Thakur and A. Chaudhary, "Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis," in *10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2022.

[12] G. A. Pal Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications,* vol. 31, p. 6863–6877, 2019.

[13] D. Rawat, "Validating and Strengthen the Prediction Performance Using Machine Learning Models and Operational Research for Lung Cancer," in *IEEE International Conference on Data Science and Information System (ICDSIS)*, Hassan, India, 2022.

[14] . L. Eldridge, "How Lung Cancer Affects Different Age Groups," 23 March2022.[Online].Available:https://www.verywellhealth.com/lung-cancer-age-5216079. [Accessed 1 January 2023].

[15] J. Gallagher, "Cancer rules rewritten by air-pollution discovery," 10 September2022[Online].Available:https://www.bbc.com/news/health-62797777. [Accessed 1 January 2023].

[16] L. Eldridge, "How Alcohol Affects Lung Cancer Risk and Outcomes," 28November2022.[Online].Available:https://www.verywellhealth.com/alcohol-and-lung-cancer-risk-2248986. [Accessed January 01 2023].

[17] "What are the Effects of Dust on the Lungs?," 3 January 2018. [Online].Available:https://www.ccohs.ca/oshanswers/chemicals/lungs_dust.html. [Accessed 01 January 2023].

[18] "Is lung cancer genetic?," WHO, 23 August , 2021. [Online]. Available:https://www.medicalnewstoday.com/articles/is-lung-cancer-genetic#summary. [Accessed 01 January 2023].

[19] "How does obesity cause cancer?," 15 June 2017. [Online]. Available: https://www.mdanderson.org/publications/focused-on-health/how-does-obesity-cause-cancer.h27Z1591413.html. [Accessed 01 January 2023].

[20] "Smoking and lung cancer: What to know and how to stop," 28 April 2021.[Online].Available:https://www.medicalnewstoday.com/articles/lung-cancer-in-non-smokers. [Accessed 01 January 2023].

[21] "Health Risks of Secondhand Smoke," 12 January 2023. [Online]. Available:https://www.cancer.org/healthy/stay-away-from tobacco/health-risks-of-tobacco/secondhand-smoke.html. [Accessed 01 January 2023].

[22] "Signs and Symptoms of Lung Cancer," 1 October 2019. [Online]. Available: https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/signs-symptoms.html. [Accessed 01 January 2023].

[23] "Fatigue experienced by lung cancer patients, in and out of treatment," 30 December 2016. [Online]. Available: https://lungcancer.net/symptoms/fatigue. [Accessed 01 January 2023].

[24] S. Felson, "How to Spot the Early Warning Signs of Cancer," 07 May 2022.[Online].Availablehttps://www.webmd.com/cancer/guide/cancer-early-warning-signs. [Accessed 01 January 2023].

[25] . C. Kuzma, "Snoring Linked to Cancer," 25 May 2012. [Online]. Available: https://www.menshealth.com/health/a19521487/snoring-linked-to-cancer/. [Accessed 01 January 2023].