International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Prediction of Lung Cancer Risk using Random Forest Algorithm Based on Kaggle Data Set

Gururaj T., Vishrutha Y. M., Uma M., Rajeshwari D., Ramya B. K.

Abstract: ***As huge amount of data accumulating currently, Challenges to draw out the required amount of data from available information is needed. Machine learning contributes to various fields. The fast-growing population caused the evolution of a wide range of diseases. This intern resulted in the need for the machine learning model that uses the patient's datasets. From different sources of datasets analysis, cancer is the most hazardous disease, it may cause the death of the forbearer. The outcome of the conducted surveys states cancer can be nearly cured in the initial stages and it may also cause the death of an affected person in later stages. One of the major types of cancer is lung cancer. It highly depends on the past data which requires detection in early stages. The recommended work is based on the machine learning algorithm for grouping the individual details into categories to predict whether they are going to expose to cancer in the early stage itself. Random forest algorithm is implemented, it results in more efficiency of 97% compare to KNN and Naive Bayes. Further, the KNN algorithm doesn't learn anything from training data but uses it for classification. Naive Bayes results in the inaccuracy of prediction. The proposed system is for predicting the chances of lung cancer by displaying three levels namely low, medium, and high. Thus, mortality rates can be reduced significantly.***

Index terms: cancer prediction, decision rules, data encoding, random forest algorithm, lung cancer, supervised learning.

I. INTRODUCTION

Across the world lung cancer is one of the major causes of the death of people. Day by day rate of lung cancer is increasing linearly concerning to the increase in population. Lung cancer is characterized by respiratory problems. The out of control cell division results in this cancer. Early detection and treatment must be provided in the initial stage so that the mortality rate can be decreased. Various body parts may also be affected by cancer due to the spreading of the affected cell. The survey conducted depicts no mechanisms is available to predict and to recognize lung cancer. There are major kinds of lung cancer i.e., SCLS and NSCLS. The most effective kind of lung cancer is NSCLS as it constitutes 84% of mortality. Lung cancer can be classified into different stages concerning to the spread of affected tissue to other glands and body parts. Due to the vast size of the lungs, infected cells are harder to find. An affected person may misinterpret cold and sourness of throat as some common infection but there is a chance of lung cancer if it lasts for a longer interval of time. Machine learning contributes to various fields. The fast-growing population caused the evolution of a wide range of diseases. This intern resulted in the need for the machine learning model that uses the patient's datasets. From different sources of dataset analysis, cancer is the most hazardous disease and it may cause the death of the forbearer.

[1] About 12% of the mortality rate across the world is due to cancer. It can be avoided and treatable in the initial stage. Most of the people are treated for the disease in later stages. This forms a major base to avoid and recognize cancer in the initial stage itself. Medical experts are treating the disease via images, identifies symptoms in later stages. Patterns of genes can be used to identify affected time in the initial stage.

[2] This disease is highly unstable with changes in the expression of genes. The practical analysis needed to know the complications and multiformity of lung cancer. Molecular-level analysis needs the study of gene patterns using microarray technology. The enormous number of genes can be examined at once using this technology

[4] The abnormal growth of cells in the body leads to the development of tumors in any part of the body. Not all tumors lead to cancer, there are two major classifications of tumors, one is a malignant tumor and the other is benign tumors. The benign tumors are non-cancerous or harmless tumors whereas the malignant tumor is harmful and leads to cancer. There are many reasons to cause cancer-based upon the genome we can classify it into two main categories such as genetic and epigenetic. The cancer is one type of a genetic disease which passes throughout the generations through heredity. This disease is characterized by cell division control failure.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
*Corresponding author:
Gururaj T., Associate Professor, Department of CSE, S J M Institute of Technology, Chitradurga, India. Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. E-mail: raj80guru@gmail.com
Vishrutha Y. M., Under Graduate Students, B.E., Department of CSE, S J M Institute of Technology, Chitradurga, India. Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. E-mail: vishrutha.ymn@gmail.com
Uma M., Under Graduate Students, B.E., Department of CSE, S J M Institute of Technology, Chitradurga, India. Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.
Rajeshwari D., Under Graduate Students, B.E., Department of CSE, S J M Institute of Technology, Chitradurga, India. Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.
Ramya B. K., Under Graduate Students, B.E., Department of CSE, S J M Institute of Technology, Chitradurga, India. Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.

The healthy cell changes into the affected one when the patterns in the cell expression i.e., cell division and growth changes. Epigenetic alterations refer to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence. The sources of cancer estimated to be 91% to 95% from habitat factors and 6% to 10% due to heredity.

[3] Accurate cancerous prediction can be accomplished only by using foretelling models for different datasets. Decision trees are coupled with the model for proper prediction. Many techniques like SVM, KNN, NB, etc can be used for nearly accurate and easy prediction. There exist particular rules for each path in decision trees.

The random forest algorithm consists of decision trees. It is one of the ensemble modes of learning. More than one tree can be constructed in this algorithm. The basis of this algorithm is weak estimators become strong when combined. Each decision trees have certain noise associated with them which is to be reduced. It produces more efficiency of 97%.

Naive Bayes classifier thinks particular character of a class is not related to any other character. It is a fast working algorithm. It has applications in real-time. More than one class prediction is possible for foretelling the chances.

The input fed to the KNN is a set of instances. These instances selected based on the nearest criteria. Also known as lazy learning because only local prediction is possible.

II. LITERATURE SURVEY

JaneeAlam, et al. [2018] have used an SVM classifier algorithm to detect lung cancer. By using an SVM classifier algorithm it can foretell the chances of lung cancer. Adjusting the digital images and partitioning is done at each stage. Adjusting digital images includes image measuring, converting an image from one color to another color, distinction improvement. Nearly 96% of cancer discovery and 86% of foretelling the chances of lung cancer is done by the present technique. By using the random forest algorithm, we can get an accurate result of up to 97%.

Abbas K. AlZubaidi, [2017] he attempted to explain and analyze the present algorithms that can group the images of cellular pathology based on their characteristics. He gave huge knowledge regarding the part of design identification in image classification. Furthermore, he introduces the latest grouping devices so these devices can recognize the content of tissue through the microscope. The outcome of the structural characteristics of the image classifier may result in false-negative. By the combination of radiology and histological images may detect suitable ways for complicated occurrences, approximate and verify entire slide imaging to build an application associated with digital pathology. By establishing the various scientific techniques lead to improve the precise results for cell identification intern it increases the duration of cancer affected people.

Raunak Dey, et al. [2018] examine the difficulties in the grouping of the benignant and malicious lung tissues in computed tomography images, which leads to study the mapping of images to target attributes. To accomplish this objective a traditional neural system is proposed, which contains a fundamental 3D Convolutional neural network – it is one of the particular kind of Artificial Neural Network based on the concept of perceptron and it falls to supervised learning that is used to data analysis, a novel multi-yield array and an increase Dense Net multi-yields. The results demonstrate decent classification performance in the lung cancer diagnosis. To overcome this automatic pulmonary nodule detection can be done, which will relax the requirement of manual annotations for nodule locations.

Saeed S. Alahmari, et al. [2018] In general, the work is based on illustrating the use of integrating delta characteristics accompanied by traditional characteristics, intern it increases the efficiency of the device by using the screening positions of the lung cancer. By using deep learning algorithm, it can discover lung cancer with the help of the combined lung screening positions.

RashmeeKohad, et al. [2015] proposed a framework that consists of 4various stages that intern discover cancer-causing lymph nodes by using various medical data set, thus it includes pre-processing, attribute removal and categorizing. The system is verified by testing the system against the 250 lung image datasets thus the strategy is applied by making use of MATLAB software (device). The SVM classifier algorithm and Artificial Neural Network algorithms are used for identifying lung cancer and the obtained results of the Ant Colony Optimization-Support Vector Machine algorithm matched with the Ant Colony Optimization-Artificial Neural Network algorithm. Ant Colony Optimization search algorithm produces the optimal results and it has high processing speed also. By using a random forest classifier algorithm will produces accurate results for the huge randomly selected dataset.

According to Arvind Kumar Tiwari, [2016] the early diagnosis of lung cancer is the toughest part because the cancer lymph nodes are structured in nature in which almost cells are crossing with one another. In order to forbid the lung cancer, the image processing approach has been used for the early diagnosis and early discovery of lung cancer and provides the remedy to the patients. In order to discover the lung cancer, the different attributes are taken out from the images and thus pattern reorganization methods are used to forbid the lung cancer. Using CT images, the SVM classification technique achieved accuracies between 78% to 98.24%. Using CT images, the Back Propagation Network classification technique achieved accuracies between 86.30% to 99.28%.

Yu. Gordienko, et al. [2017] has done research work on a deep learning algorithm. The effectiveness of lung segmentation and bone shadow ejection methods for examining the 2D CXRs by applying the deep learning algorithm, it guides the radiologists in order to find out the doubtful lesions and nodules of cancer affected people. It highlighted the pre-processed dataset without bones are larger precision than pre-processed dataset after lung division and avoids the resulting loss and produces greater efficiency. The training itself increases the capacity and difficulties of the deep learning algorithm having a reduced length of seven layers varies to greater than a hundred layers as same as the present networks.

Training the much complex method over the larger images and considering the fine-tuning datasets along with deep learning methods, thus it acts as a critical part for the accuracy of the methods being used.

Supreet Kaur, et al. [2016] have researched by making use of possible categorization based on data mining algorithms such as Bacterial Foraging Optimization, Linear Discriminant Analysis and Neural Network having an enormous capacity of healthcare-related datasets. By accumulating a large volume of healthcare-related data, regrettably, that is not mined for detecting the invisible facts. By taking a general lung cancer sign, includes breathing problems, age, gender, smoke, dust allergy, shoulder pain thus it can discover the probability chances of getting lung cancer. In the existing system drawing out the medical datasets will only precis for the on-going attempt. The main idea behind this is to add the hybrid grouping plan that makes the data mining devices suitable for the medical diagnosing center. In order to overcome this issue, the random forest algorithm is used to train general lung cancer datasets.

III. PROPOSED SYSTEM

A. Problem Statement

Currently, detection of cancer is performed by experienced medical professionals; professionals must often search through many similar cases before finding the desired label or class of the patient. This process of manual recognition is slow and possesses a degree of subjectivity which is hard to be quantified. To design and implement an efficient and accurate Machine learning platform that can help to predict the possibilities of early-stage lung cancer.

The input given to the model is twenty-three different features that are essential to a lung cancer diagnosis. i.e.,

Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoke, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing

The processing methods used are standardization and label encoding. The random decision forest algorithm is used as a Classifier.

Result obtained will be the Probability of having Lung Cancer.

B. System Architecture

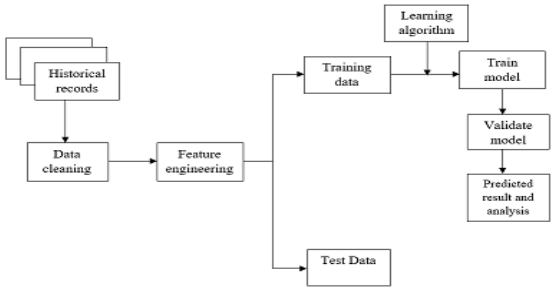

Fig.1 Architecture of the proposed system.

To design a web application, the architecture of a system forms the base. The system architecture is illustrated in Fig.1. The data is collected from the previous records of the affected people. Only the useful data is filtered from the previous records and the cleaned data is been extracted. By using the feature engineering technique, the data has been split into training data and test data. Test data is for testing the model. Training data make use of the learning algorithm to train the model. The model which is trained gets verified from the trained data. Then the model foretells the result and gets evaluated.

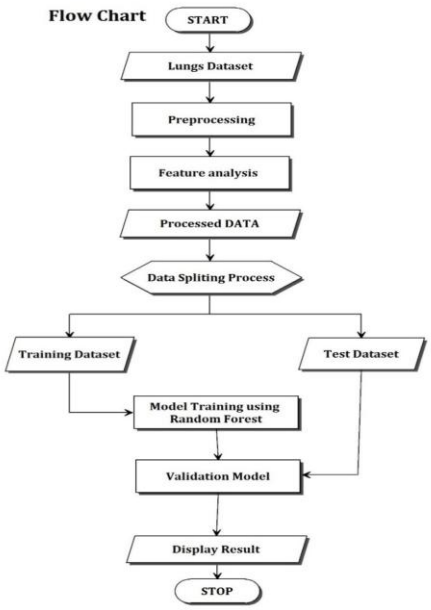

Fig.2 Flow diagram of Random forest technique.

It is necessary to conclude all the functions and reach the limits. The flow diagram acts as a foundation for system design. In order to run the function and for the scheduling process we use a flowchart. The main advantage of using the flowchart is that it displays all the tasks that are mixed up in a system within the straight value chain. The prediction of the disease is depicted in the flowchart. The flowchart shows the series of steps that depicts one or more inputs and modifies them to outputs. Flowchart of our project is illustrated in Fig.2. The lungs dataset is taken from the past data of the affected people. Using the feature analysis, the data is been processed. The processed data is split into training data and test data by making use of a data splitting process. The training dataset trains the model using the random forest algorithm. The test dataset is used to validate the model. Finally, it displays the result whether it is low, medium or high.

IV. DATA COLLECTION AND PROCESSING

Datasets are collected from free online repositories like UCI, Kaggle.com, etc. These datasets are in numerical format.

It contains 23 columns with unique numerical values. Datasets fed as input to the model responsible for the model learning process.

A dataset with 23 columns collected:
Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoke, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing.

A. Phase I

- Data Acquisition and Pre-processing
- Feature Selection and Data Preparation

Data Acquisition and Pre-processing

The basis for machine learning is data and models. During the collection of data, make sure that the collected data have enough features to train the model. The data that are collected from online sources may contain statements, numbers, and approximate terms. The basic form of data contains mistakes, exclusion, inconsistencies. The steps to process this primary data are, a large amount of fresh data is gathered from the research and are classified into individual group with respect to related factors of independent responses. Data pre-processing technique is more important due to the existence of unwanted data.
Real-time data is made up of-
Inaccurate data - Inaccuracy in data may happen for several reasons such as data collected may not be in continuous form, misplacing of data during entry or else because of some technical problems and some more.
Noisy data (bogus and irregularity) - Large amounts of meaningless data may be collected due to the technical issues in apparatus which assemble data, user-caused errors while recording of information.
Inconsistent data - Inconsistency in data is because of the presence of replication of data, human mistakes during data entry, holding errors in programs.

Feature selection and preparation

Feature engineering is a discipline that makes data to generate some properties which are responsible for the working of ML algorithms. If this process is carried out precisely, then the machine learning algorithm increases its forecasting power by generating some properties from fresh data which helps in lubricating the machine learning techniques.
Feature engineering is a great facility in machine learning which generates a large gap in the middle of a good and a bad model. It is a technique that reshapes raw data into properties that symbolizes intrinsic problems to the output models.
Classification means the process that divides data into groups based on some aspects. So data visualization is to be made, to see whether the training data have a correct label, which is called a target value.
The next process is dividing the data as training data, test data. Here training data is subdivision to train a model and the test set is a subdivision to test the trained model.
The test set should satisfy the following two constraints they are, Train set should be large enough to yield meaningful results. The next one is the test set selected should not have other characteristics than the train set.

B. Phase II

- Model implementation and learning
- Model testing and output

Model implementation and learning

Model implementation refers to feeding the specific machine learning algorithm along with the training datasets into the model. Thus, the model learns from those. The machine learning model is nothing but a piece of code that forms the core of the system and responsible for the processing. The accurate training datasets should be used for building the proper model. The machine learning algorithm is used for grouping the individual details into categories to predict whether they are going to expose to cancer in the early stage itself. Random forest algorithm implemented. It produces more efficiency of 97% comparatively. The multiple trees for decision making are constructed automatically by the algorithm. Each tree outputs the result called the vote. The majority voting process is used to aggregate individual votes from trees to form the final predicted value. First, take out the test characteristics of each and every automatically created trees to foretell the outcome and preserve the forecasted target value. Secondly, it computes the votes generated for each forecasted outcome. Finally, examine the large voted forecast value as the terminated target value obtained from the algorithm.
This algorithm is implemented to carry out the process of foretelling we must pass the test characteristics across the order of each and every randomly generated tree.
Consider there are 100 decision trees obtained randomly from the random decision forest algorithm. Every decision forest can foretell incompatible outcomes for identical test characteristics. Then compute each outcome votes. Let us consider 100 decision trees, it generates a distinctive outcome such as “a, b, c” in which ‘a’ must be nothing but out of 100 trees how many trees will predict ‘a’, similarly for the other two outcomes (b, c). In case a gets the highest precision than out of 100 trees 60 trees are forecasted as the outcome is ‘a’. Thus, random decision forest results ‘a’ will be the target value. The overall idea is known as majority voting.

Model testing and output

The testing of the model includes giving unique datasets. Datasets are divided into test data and train data in the earlier phase itself. The outcome is to obtain a proper accurate model for prediction. The output of the model is predicting the chances of lung cancer by displaying three levels namely low, medium, and high. Thus, mortality rates can be reduced significantly. Test data used should be randomly given while testing and high accuracy of the model during the testing phase might indicate that test data has leaked into the train data.

V. IMPLEMENTATION

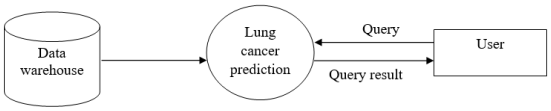

Fig.4.1 Level-0 model for Lung cancer prediction.

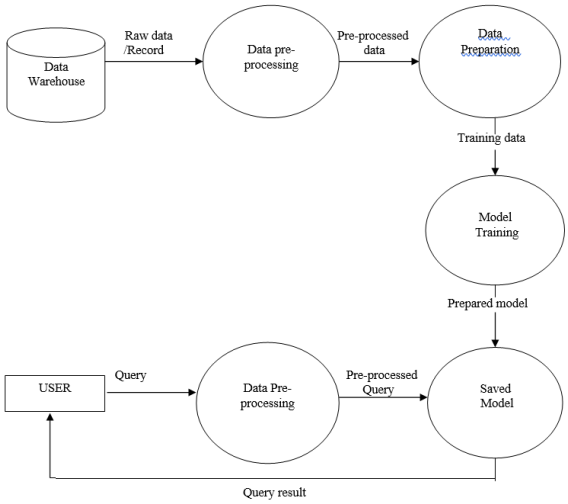

Fig.4.2 Level-1model for lung cancer prediction.

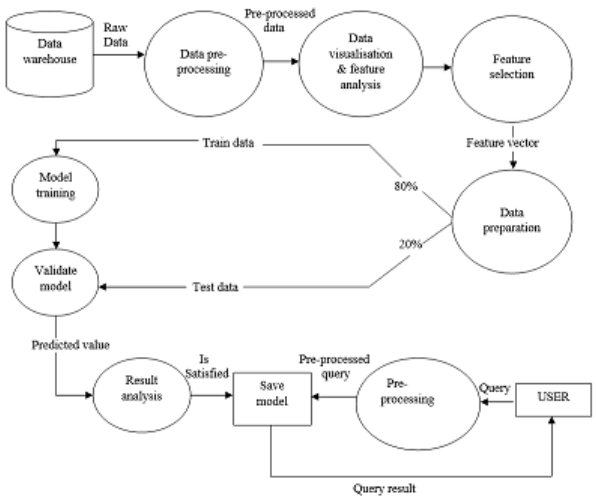

Fig.4.3 Level-2 model for lung cancer prediction.

Implementation of the model is depicted above. These diagrams show the levels of the model. In level 0, the model collects the data from the data warehouse. The user can send the queries to the model that foretells the results as shown in fig.4.1. In level 1, the fresh data is collected from the warehouse. The pre-processed data is fed into the data preparation and thus the training data is obtained. The saved model foretells the results to the user as shown in the fig.4.2. The user queries are pre-processed and saved in the model. Both the levels 0 and 1 are combined to form the level2.The splitting of cleaned data into 0.80 and 0.20 percent. The split data used for both training and validating purpose. The predicted value is sent to the result analysis and if the predicted value is satisfied then it is saved as depicted in fig.4.3.

VI. RANDOM FOREST ALGORITHM

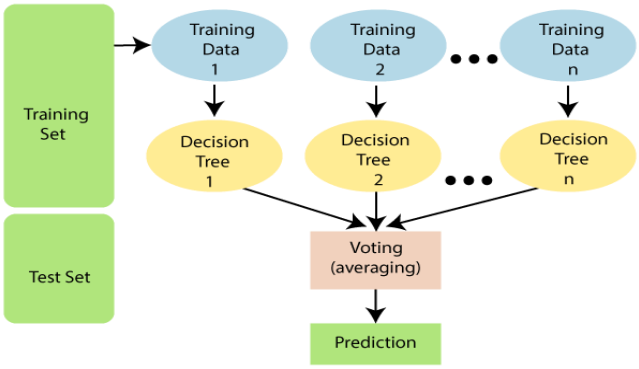

Fig.5 Design of random forest algorithm.

A.Random Decision Forest Algorithm

Input: Samples are fed as input from dataset.
Output: Final prediction value is obtained as low, medium, high.

Pre-Condition: A training set T= (a1, b1), characteristics C, number of trees in forest D
Step 1: fun RandomDecisionForest (T, C)
Step 2: G ←0
Step 3: for O∈1…, D do
Step 4: T (o)← bootstrap sample from T
Step 5: go ←RandomizedTreeLearn (T (e), c)
Step 6: G←G ∪ {go}
Step 7: end for
Step 8: return G
Step 9: end fun
Step 10: fun RandomizedTreeLearn (T, C)
Step 11: at each node:
Step 12: c← very small subset of C
Step 13: split on best feature in c
Step 14: return the learned tree
Step 15: end fun

B. Working of Random decision forest result prediction pseudo code is depicted below

First, take out the test characteristics of each and every automatically created trees to forecast the outcome and preserve the forecasted target value. Secondly, it computes the votes generated for each forecasted outcome. Finally, examine the large voted forecast value as the terminated target value obtained from this algorithm. The design of the random decision forest algorithm is illustrated in fig.5.
This algorithm is implemented to carry out the foretelling process. The test characteristics are transmitted across the order of each and every randomly generated tree.

Consider there are 100 decision trees obtained randomly from the random decision forest algorithm. Every decision forest can forecast incompatible outcomes for identical test characteristics. Then compute each outcome votes.

Let us consider 100 decision trees, it generates a distinctive outcome such as "a, b, c" in which 'a' must be nothing but out of 100 trees how many trees will predict 'a', similarly for the other two outcomes (b, c). In case a gets the highest precision than out of 100 trees 60 trees are forecasted as the outcome is 'a'. Thus, random decision forest results 'a' will be the target value. The overall idea is known as majority voting. The fig.6 depicts the flow of the algorithm.

C. Random Forest Working

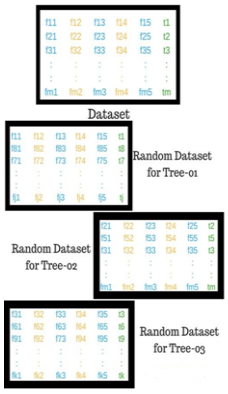

Fig.6 Working of the random forest algorithm.

D. Merits of Random Decision Forest Algorithm

Some of the merits of this algorithm are depicted below:

1. The predictive performance can compete with the best-supervised learning algorithm.
2. They provide a reliable feature importance estimate.
3. The random decision algorithm used for categorization and regression analysis tasks.

VII. RESULTS AND DISCUSSIONS

A. Heatmap Of Data Visualization

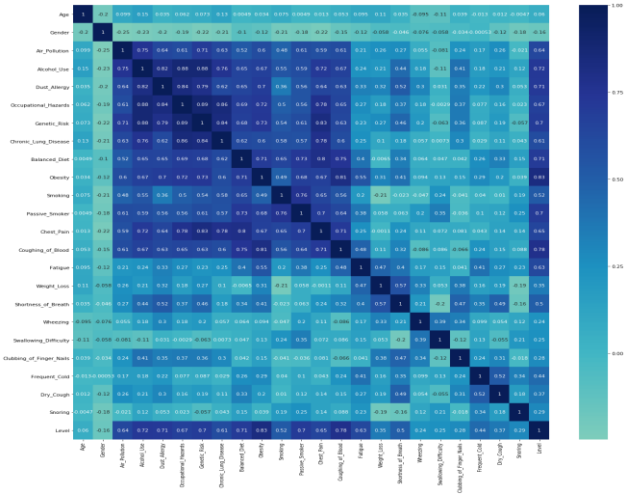

Fig.7 Graphical representation of heat map.

Heat maps are the graphical depiction of the features and it uses a color-coded framework for its depiction as shown in fig.7. The main cause of using heat maps is to better envisage the capacity of the events in the datasets. Heat maps play a major role in visualizing the useful data. It helps the observer to analyze the data more easily. It uses a dark-to-light color scale.

B. Classification Report

This is implemented to evaluate the quality of predictions from a random forest algorithm. It depicts the precision, recall, f1-score and support scores for the model as shown in table1.

Table1: Classification report

	Precision	Recall	f1-score	Support
0	1.00	1.00	1.00	55
1	1.00	1.00	1.00	63
2	1.00	1.00	1.00	82
Micro avg	1.00	1.00	1.00	200
Macro avg	1.00	1.00	1.00	200
Weighted avg	1.00	1.00	1.00	200

C. Summarizing The Concept Of Classification Report

Table2: Summarizing the concept of classification report

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive – it is the result in which the model perfectly foretells the positive class.

True Negative – it is the result in which the model perfectly foretells the negative class.

False Positive – it is the result in which the model imperfectly foretells the positive class.

False Negative – it is the result in which the model imperfectly foretells the negative class.

The overall classification report is summarized in the table2.

D. Precision

The classifier can label a sample as a positive and a negative. The values range between 1 and 0 respectively. The total predicted positive value is shown in the table3.

Table3: Total predicted positive value.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Positive = Total Predicted Positive

FORMULAE:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{\text{True Positive}}{\text{Total Predicted Positive}} \quad (2)$$

E. Recall

The classifier can discover all the samples which are positive as depicted in the table4. The best value and the worst value for the recall are 1 and 0 respectively.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Negative = Actual Positive

FORMULA:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{\text{True Positive}}{\text{Total Actual Positive}} \quad (3)$$

F. F1 Score

It can be simplified as a loaded median of precision and recall. The values range between 1 and 0 respectively.

FORMULA:

$$F1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

G. Confusion Matrix for Model Validation

A confusion matrix can be represented in a matrix form as depicted in fig.8. This matrix is for calculating the efficiency of a supervised machine learning algorithm. The instances of the actual and the predicted classes are depicted by each elements of the confusion matrix.

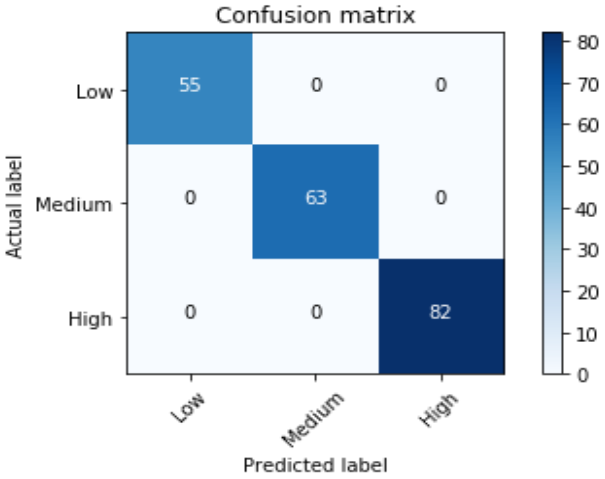

Fig. 8 Confusion matrix for model validation

H. Comparison of Algorithms

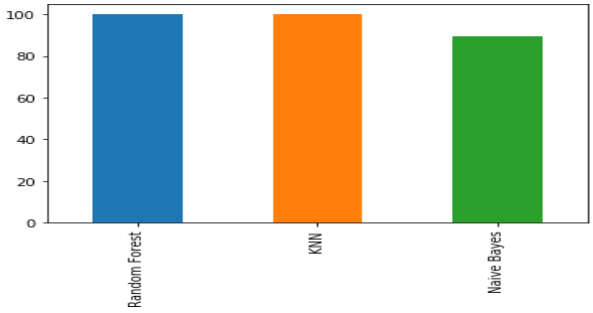

Fig.9 Comparison of algorithms using accuracy graph.

Random forest algorithm is also known as random decision forests. The algorithm is used for learning the technique for categorization, reverting and function that run by building a mass of decision tree training schedule and outputs the class which is the system of classes. It gives the most precise output. Weak estimators combine to form strong estimators. This is the principle on which the random forest algorithm works. The output of the model will be proper even if there are vulnerable inputs from one or a few decision trees. If the number of estimators is small, approximately 30 then it produces high accuracy which is equal to 97%.

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The accuracy of the NB classifier is less compared to the Random forest and KNN.

The k-nearest neighbors'-algorithm (KNN) is a non-parametric method used for classification and regression. It has equal efficiency as Random forest. A greater accuracy compared to NB classifier can be noticed. But KNN prediction results in delay with respect to time.

By comparing random forest, KNN, and Naive Bayes algorithms, the accurate result is obtained from the random forest algorithm. Using the accuracy graph the comparison of algorithms is shown in fig.9.

VII. CONCLUSION

Now a day's lung cancer is also considered as the widely spreading disease. The process of manual recognition is slow and possesses a degree of subjectivity which is hard to be quantified. This paper deals with design and implementation of an efficient and accurate Machine learning platform that can help to predict the possibilities of early-stage lung cancer. The future enhancements of this system can be made to obtain more accurate model from KNN and Naive Bayes along with the Random forest algorithm. The automation of lung cancer can be made by utilizing datasets directly from the hospitals and various agencies rather than repositories. The algorithms like CNN, ANN, AI algorithms can be implemented.

VIII. ACKNOWLEDGEMENT

This research was supported by SJM Institute of Technology, Chitradurga, and Visvesvaraya Technological University, Jnana Sangama, Belagavi -590018.

REFERENCES

1. L. Shoon et al., ''Cancer recognition from DNA microarray gene expression data using averaged one-dependence estimators''.
2. G. Russo, C. Zegar, and A. Giordano, ``Advantages and limitations of microarray technology in human cancer".
3. X. Wang and O. Gotoh, ''Microarray-based cancer prediction using soft computing approach''.
4. Bashetha and G. U. Srikanth, ''Effective cancer detection using soft computing technique''.
5. Janee Alam1, Sabrina Alam and Alamgir Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier" on 12 March 2019.
6. Abbas K. AlZubaidi, Fahad B. Sideseq, Ahmed Faeqand Mena Basil , "Computer Aided Diagnosis in Digital Pathology Application: Review and Perspective Approach in Lung Cancer Classification" , Annual Conference on New Trends in Information & Communications Technology Applications-(NTICT'2017) 7 - 9 March 2017.
7. Raunak Dey, Zhongjie Lu and Yi Hong, "Diagnostic Classification Of Lung Nodules Using 3D Neural Networks",Accepted for publication in IEEE International Symposium on Biomedical Imaging (ISBI) 2018.
8. Saeed S. Alahmari , Dmitry Cherezov, Dmitry B. Goldgof , (Fellow, Ieee), Lawrence O. Hall, (Fellow, Ieee), Robert J. Gillies And Matthew B. Schabath , "Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening", Received October 30, 2018, accepted November 13, 2018, date of publication November 29, 2018, date of current version December 31, 2018.
9. Arvind Kumar Tiwari, "Prediction Of Lung Cancer Using Image Processing Techniques: A Review", Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.1, January 2016.
10. Yu.Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu.Kochura, O.Alienin, O. Rokovyi, and S. Stirenko, "Deep Learning with Lung Segmentation and Bone Shadow Exclusion Techniques for Chest X-Ray Analysis of Lung Cancer", on 2017.
11. Supreet Kaur and Amanjot Kaur Grewal, "A Review Paper on Data Mining Classification Techniques For Detection Of Lung Cancer", International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 11 Nov -2016.
12. RashmeeKohad and Vijaya Ahire, "Application of Machine Learning Techniques for the Diagnosis of Lung Cancer with ANT Colony Optimization", International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015.

AUTHORS PROFILE

Gururaj T., is a Research Scholar at Ramaiah Institute of Technology(MSRIT), Bangalore affiliated to VTU. He received his Master of Technology in Computer Science & Engineering from J. N. National College of Engineering, Shivamogga. (VTU). He is a part-time research scholar at MS Ramaiah Institute of Technology, Bangalore and currently working as an Associate Professor in the Department of Computer Science and Engineering at S. J. M. Institute of Technology, Chitradurga. His main area of interest includes studies related to big data and its applications and Bioinformatics.

Vishrutha Y. M., is an undergraduate student in Computer Science stream at SJM Institute of Technology and will be graduating in 2020 with BE in Computer Science.

Uma M., is an undergraduate student in Computer Science stream at SJM Institute of Technology and will be graduating in 2020 with BE in Computer Science.

Rajeshwari D., is an undergraduate student in Computer Science stream at SJM Institute of Technology and will be graduating in 2020 with BE in Computer Science.

Ramya B. K., is an undergraduate student in Computer Science stream at SJM Institute of Technology and will be graduating in 2020 with BE in Computer Science.