International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

Early Stage Prediction of Lung Cancer Using Machine Learning

Bhavya M R[1], Rajat Govil R[2], Sandesh Giri B C[3], Varshith B R[4], Veerabhadra Swamy N S[5]

Assistant Professor, Department of CS &E, Maharaja Institute of Technology, Mysore, India [1]

UG Student, Department of CS &E, Maharaja Institute of Technology, Mysore, India[2]

UG Student, Department of CS &E, Maharaja Institute of Technology, Mysore, India[3]

UG Student, Department of CS &E, Maharaja Institute of Technology, Mysore, India[4]

UG Student, Department of CS &E, Maharaja Institute of Technology, Mysore, India[5]

ABSTRACT: The prediction of patients prone to lung cancer can help doctors in their decision making regarding their treatments. So that the doctor can get to know whether the patient is suffering from lung cancer, he can start the treatment at early stage will be more beneficial for the patient, were he can diagnosed and chances of getting rid of lung cancer is more. In this regard, this system attempts to evaluate the discriminative power of several predictors in the study to increase the efficiency of lung cancer detection through their symptoms. ML Classifier Naïve Bayes (NB) is evaluated on a benchmark dataset obtained from UCI repository to detect lung cancer. We take different attributes such as smoking rate, blood Pressure, age, weight etc. of various patients who were already predicted the result. The new patients same attributes are considered to give as input for the trained model for the prediction. Our model will predict the patient is prone to lung cancer or not. This result is more accurate compare to manual analysis by the doctor this gives much more accuracy to doctors for lung cancer prediction**.**

KEYWORDS: Healthcare, Classification, Navie Bayes, Lung Cancer

I. INTRODUCTION

Lung cancer is a condition that causes cells to divide in the lungs uncontrollably. This causes the growth of tumours that reduce a person's ability to breathe. Lung disease is standout amongst the most widely recognized malignancies, representing more than 225,000 people, 150,000 deaths, $12 billion cost yearly. It is as well one of the deadliest diseases; just 17% of individuals determined to have lung tumour survive 5 years after identification, survival rate is bringing down in developing nations. The most important cause of lung cancer is smoking (cigarette, pipe, and cigar). It has been found that 90% of all lung cancer cases have been due to tobacco smoking. Tobacco smoke contains around 4000 chemicals, out of which about 60 are found to be carcinogenic Other causes of lung cancer are, Radon(Naturally occurring gas),Asbestos(Natural mineral used in construction),Metals such as cadmium, chromium, arsenic exhausted from diesel engine. Chemicals in and certain diseases that affect the lung (e.g. tuberculosis).Family history of cancer can also put a person at greater risk. The closer the relative, the greater is the risk.

. Comparing with the other cancer the prediction of the lung cancer at the early stages is a quite challenging task to the doctors Also it requires a lot of experience for a doctor but also with the less accuracy. The doctor need more experience to predict the cancer at the early stages although he is experienced sometimes he may fail to predict the cancer at early stages.Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

II. RELATED WORK

Muhammad Imran Faisal., [1] claimed that even lung cancer can be predicted in early stages without any instrumental help. The prediction of patients prone to lung cancer can help doctors in their decision making regarding their treatments. In this regard, this research paper attempts to evaluate the discriminative power of several predictors in the study to increase the efficiency of lung cancer detection through their symptoms. A number of classifiers including Support Vector Machine (SVM), C4.5 Decision tree, Multi-Layer Perceptron, Neural Network, and Naïve Bayes (NB) are evaluated on a benchmark dataset obtained from UCI repository. The performance is also compared with well-

known ensembles Such as Random Forest and Majority Voting. Based on performance evaluations, it is observed that Gradient-boosted Tree outperformed all other individual as well as ensemble classifiers and achieved 90% accuracy.

Moffy Vas, Amita Dessai., [2] suggested that image processing tools can be used for early detection of cancer. In this paper, a lung cancer detection algorithm is proposed using mathematical morphological operations for segmentation of the lung region of interest, from which Haralick features are extracted and used for classification of cancer by artificial neural networks.

Bipin Nair B.J ., [3] suggested that Cigarette smoking has a significant role in changing MiRNA expression. There are many miRNAs which have significant role in the intensity of airway blockade in chronic obstructive pulmonary disease (COPD). Our objective is to Measure the possibilities of occurring cancer. To reach our goal we follow three steps. First one is predicting the secondary structure of both normal LC-miRNA and defective LC-miRNA. In comparing both the secondary structures defective region is extracted. The second step deals with predicting the target region of normal LC-miRNA using Target. Scan, an online software. The third step involves comparing the binding target of normal LC-miRNA and defective LC-miRNA to generate an optimal result. This will simplify cancer detection procedures and possibly would help to prevent cancer for people with lung cancer genes. As a result of this research 50% of miRNA targets were predicted accurately.

Jane Alam., [4] suggested that Depression is one of the most common mental disorders affecting millions of people worldwide. Developing adjunct tools aiding depression assessment is expected to impact overall health outcomes and treatment cost reduction. There have been reported researches for cancer cell detection in recent year. Murphy *et al.* built up a CAD framework, where lungs pictures were divided by utilizing the region growing technique and morphological smoothing. The algorithm had an accuracy of 84%.Proposed another algorithm to improve the location of knobs with ground-glass opacity. Messy, Hardier and Rogers displayed a CAD algorithm utilizing thresholding, morphological handling and Fisher Linear Discriminant to fragment, recognize patient's nodules and take out of false positives. The framework got an accuracy of 82.66% with 3 FP per case being validated with 143 knobs. Gomati and Thangaraj utilized image processing algorithm, Fuzzy CMean calculation and neural classifier in the phases of pre-processing, fragmentation identify patients nodules and respectively. This algorithm had an accuracy of 76.9%.Kumar roposed a CAD algorithm that utilized Biorthogonal Wavelet Transform, region growing and fuzzy based framework in pre-processing, fragmentation and identification of nodules. The algorithm had an accuracy of 86%.

Sanjukta Rani Jena., [5] reviews the research of diagnosing Lung cancer using ML algorithm and suggests how ML techniques can be employed and worked in practice. According to the paper, Traditional ML algorithms such as Support Vector Machines (SVM), Gradient Boosting Machine (GBM), Random Forest, Naïve Bayes, and K-Nearest Neighborhood (KNN) are frequently used for Lung Cancer area researches were systematically organized and summarized. The paper suggested that the Researchers using ML algorithms should be aware of the properties of their ML algorithms and the limitation of the results they obtained under restricted data condition.

III. SYSTEM DESIGN

Systems design is a process in which it is used to defining the interfaces data and modules for a system to specify the requirements. The main purpose of the system design is to define architecture of the system, logical data flow of the system, physical design of the system. The purpose of the system design is to develop the system architecture by giving the data and information that is necessary for the implementation of a system.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

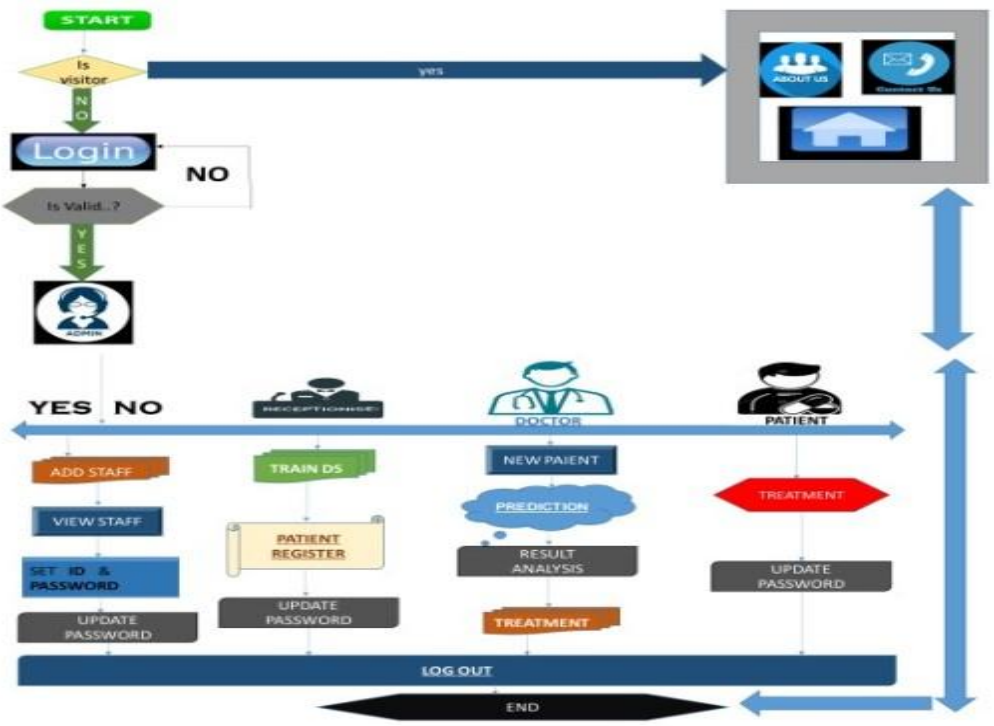

Fig 1: Flow Chart

IV. METHODOLOGY

ML concerns with construction and study of system that can learn from data. For example, ML can be used in E-mail message to learn how to distinguish between spam and inbox messages.
There are there types of Machine learning (ML), they are
i. Supervised Machine Learning
Here we have labels and the input is past examples.
ii. Unsupervised Machine Learning
Extraction of patterns without labels.
iii. Semi-Supervised Machine Learning

Mixture of both Supervised and Unsupervised Machine Learning
In the project we use supervised learning techniques to process medical data-set. We use Naive Bayes Algorithm to predict lung cancer.
Reasons for selecting Naive Bayes;
1. Most of the previous medical research papers uses this algorithms.
2. Survey says efficient algorithm to process medical data.
3. Takes less time for data processing.
4. Works fine for n number of parameters. Number of parameters need not to be fixed.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

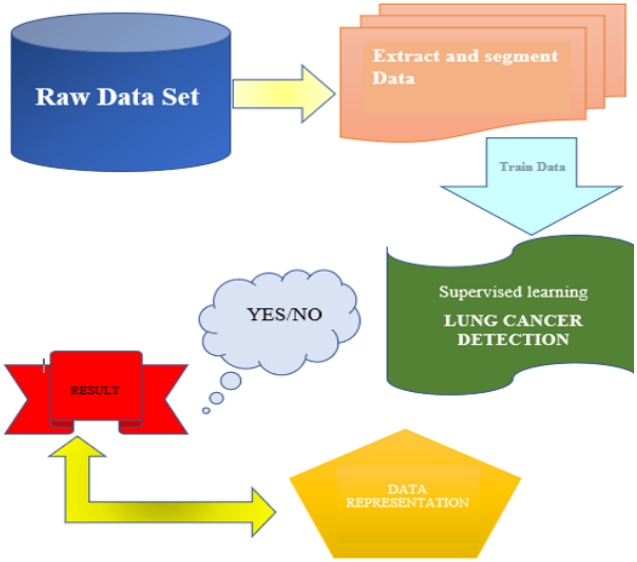

Fig 2: Methodology

1. Raw Dataset:

Data is a collection global dataset. In this system use UCI data set is used for training a model. Data set contain 8 parameters and around 1000 dataset. The dataset feature/parameters are:

- Age
- Gender
- Weight
- Blood pressure
- Sugar tested value
- Smoking
- Family history
- Alcoholic.

This are data is trained to the model for the prediction of lung cancer.

2. Train Dataset and Test Dataset:

The training data is an initial set of data which is used to understand the program. This is the one in which we have to train the model first because to set the feature and this data is available on system. This data is used to teach the machine for do different actions. It is the data in which model can learn with algorithm to teach the model and doing work automatic.

Testing data is the input given to a software. It shows the data affects when the execution of the module that specifying and this is basically used for testing.

3. Pre-processing of data

Data pre-processing is a process in which that is actual use for converting the basic data into the clean data set. It is the step in which the data transform or an encode to the state that the machine can be easily parse. The major task of data pre-processing in learning process is to remove the unwanted data and filling the missed value. So that it help to machine can be trained easily.

4.Feature Extraction

Feature Extraction is the method in which it used for alter the key data for features of outcomes. This, trait square is used to compute the characteristics of designs given that facilitate in different amid the class of key pattern details. This method involving to reducing the counts of resource required to describe the huge set of data. Feature extraction is an attribute reduction process. This is also used to increasing the speed and effectiveness of supervised learning.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

5. Machine Learning Algorithm:

Naïve Bayes Algorithm:-The Navie Bayes is a ML algorithm is the non-parametric method proposed by Thomas Bayes used for Classification. It is a technique for classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set The input has numeric values so that we can check the probability of each attribute. Navie Bayes algorithm is a type of instance based learning. This algorithm relies on the distance for classification the training data can improve its accuracy dramatically. The attributes are taken from a set of objects for which the class or the object property value is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

6. Result:

After taking that input data from the system will able to divine the statistics by appeal the ML algorithm & also provide the foremost output in the device. Which predicts whether a person is suffering from lung cancer or not. And it will provide Treatment along with precautions

V. RESULT AND DISCUSSION

In our project the result is classified into Yes or No. If the result is classified into yes then the patient is suffering from lung cancer. If the result is classified into No then the patient is not suffering from lung cancer. We analyze the result of the lung cancer prediction and check the accuracy of the lung cancer prediction, time taken to compute the accuracy of the diabetes prediction, correctly classification and incorrectly classification of result of the lung cancer prediction. We have used Naïve bayes Algorithm to predict the lung cancer where result is classified into Yes or No. We compared the testing data and actual data to get the accuracy of our project.

Constraint	Naive Bayes
Accuracy	100%
Time(mille seconds)	1481
Correctly Classified	100%
Incorrectly Classified	0%

VI. CONCLUSION AND FUTURE SCOPE

Building lung cancer prediction system is useful for hospitals and doctors. System predicts lung cancer at early stages, so doctors can treat patients in a better way. Proposed system is an real time application which is meant for multiple hospitals and predicts lung cancer in less time. As we use machine learning algorithms for lung cancer prediction, we will get more accurate and efficient results. It is successfully accomplished by applying the classification algorithms. This classification technique comes under data science technology.

Future Scope: This System is built for early stage prediction of lung cancer further analysis can be made for prediction of lung cancer at several stages

SMS/Email Module – In the proposed system, admin assigns Id and password for doctors and receptionists and is intimated manually, so we can add SMS/Email module as a future enhancement where doctors and receptionists receives an SMS or Email regarding the Id and password.

Query Module- we can add the query module as a future enhancement to the application where doctor, receptionist and admin of the application can interact with each other

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

REFERENCES

[1] Cabrera, J., Dionisio, A., & Solano, G. (2015, July). Lung cancer classification tool using microarray data and support vector machines. In Information, Intelligence, Systems, and Applications (IISA), 2015

[2] Yu, Z., Chen, X. Z., Cui, L. H., Si, H. Z., Lu, H. J., & Liu, S. H. (2014). Prediction of lung cancer based on serum biomarkers by gene expression programming methods. Asian Pacific Journal of Cancer Prevention, 15(21), 9367-9373.

[3] Wender, R., Sharpe, K. B., Westmaas, J. L., & Patel, A. V. (2016). The American Cancer Society's approach to addressing the cancer burden in the LGBT community, 3(1), 15-18.

[4] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., & Cui, Q. (2013). HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic acids research, 42(D1), D1070-D1074.

[5] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[6] Wender, R., Sharpe, K. B., Westmaas, J. L., & Patel, A. V. (2016). The American Cancer Society's approach to addressing the cancer burden in the LGBT community. LGBT health, 3(1), 15-18.

[7] Hussain. S, Keung. J, Khan. A. A., Performance evaluation of ensemble methods for software fault prediction, An experiment, Proceeding of ASWEC, 2015.

[8] Hussain. S, Asghar. Z, Ahmad. B, Ahmad. S., A step towards software corrective maintenance using RCM model, International Journal of Computer Science and Information Security, 4(1), 2009.

[9] Karakach, T. K., Flight, R. M., Douglas, S. E., & Wentzell, P. D. (2010). An introduction to DNA microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems, 104(1), 28-52.

[10] Marusyk, A., Almendro, V., & Polyak, K. (2012). Intratumour heterogeneity: a lookingglass for cancer?. Nature Reviews-Cancer, 12(5),323