IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 12, December 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

Automated Disease Analyzer and Post Operative Life Anticipation after Thoracic Surgical Procedures

Avni Goel [1], Iksha Joshi [2], Isha Singh [3]

U.G. Student, Department of Information Technology, HMR Institute of Technology and Management, New Delhi, India[1]

U.G. Student, Department of Information Technology, HMR Institute of Technology and Management, New Delhi, India[2]

Associate Professor, Department of Information Technology, HMR Institute of Technology and Management, New Delhi, India[3]

ABSTRACT: Keeping a track of health-related conditions to lead a normal life is important. Taking necessary steps by keeping track of your health ensures high quality of livelihood and also helps healthcare workers. Thoracic surgeries collectively refer to surgical operations that are organs of the chest area. Health issues are often seen post thoracic surgical procedures which are a huge concern among healthcare workers. Moreover if one is able to analyze and predict diseases at the initial phase these types of chronic illnesses can be detected.Little research and recommendation models have been seen for predicting diseases at initial phases as well as life anticipation post lung cancer surgeries. With the use of machine learning techniques, we aim to effectively predict diseases at initial phases as well postoperative mortality after 1 year in patients using computational methods. The success rate and efficiency can be depicted from the results achieved.

KEYWORDS: Lung cancer, Thoracic surgeries, anticipation

I. INTRODUCTION

Due to lifestyle changes and environmental factors, we have seen a recent surge in diseases in people. Due to this people often suffer from Chronic diseases and one of them is lung cancer. The major concern among healthcare workers is the mortality rate of patients after thoracic surgeries one of which is lung cancer. The first year after Thoracic surgery of lung cancer is very crucial for patients, so it becomes an important task for doctors to ensure that patients live a normal life post thoracic surgery. In Today's situation , selective methods are present to handle risks or complications after performing thoracic surgery.

Large Datasets often affect the overall accuracy of Machine learning algorithms which in return also affects the computing time of anticipation. To solve this problem several selection methods are there so that attributes that are inconsequential can be removed because it directly affects the processing of data. The project aims to predict diseases at initial phases which will prove to be a vital factor to prevent many diseases. Also, if a person is detected with lung cancer and has undergone thoracic surgery, an attempt is made to check life anticipation after 1 year. In Future doctors might be able to enhance their treatment techniques and may adopt more advanced medication to treat patients that will help them to live a normal life.. Best possible machine learning techniques can be compared and analysed to do so like using naive Bayes, logistic regression, and random forest ML algorithms to get maximum accuracy to achieve the target of detecting diseases at initial phases as well as the mortality rate of patients after Thoracic Surgical procedures.

II. RELATED WORK

Thoracic surgeries done on lung cancer patients is a very vital factor because it tells the post operative Life Anticipation of patients after one year. Furthermore if diseases are detected at early stages,fatal diseases like lung

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

cancer can be prevented. Various related works are published for life anticipation of post and pre operative thoracic surgeries.

Abeer S. Desuky and Lamiaa M. el Bakrawy [7] have used attribute ranking and selection methods in machine learning techniques to anticipate life anticipation post thoracic surgeries. To do so, a WEKA tool was used. Machine learning algorithms such as Naive Bayes, logistic regression, Multilayer Perceptron and J48 that were taken into consideration were implemented and performed by doing required coding. It was further compared to their boosted versions to map the efficiency . The results showed that boosted versions did not give much accuracy as compared to attribute ranking and selection methods.

Adam Abdulhamid, Ivaylo Bahtchevanov, Peng Jia[3] implemented a feature set by classifying data in accordance with health issues after Thoracic surgery. They used a dataset of 470 patients each having 16 variables. Machine Learning techniques like Naive Bayes, Random forest were compared and the results were compared to anticipate whether the patient lived after Thoracic surgeries.

It was seen that Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma[4] used linear regression as the the best machine learning technique to predict life expectancy for various diseases as well as predicting which disease was more prominent in various continents.

III. METHODOLOGY

The suggested methodology aims to give an easy understanding, consisting of sequential steps for the approach. To calculate both life anticipation post thoracic surgery and early detection of various diseases at initial phases this project is split into 2 parts :

A. Prediction of life post thoracic surgery within 1 year.
B. Automated disease Analyzer at initial phases

A. Prediction of life post thoracic surgery within 1 year

From UCI Repository Statistics are taken of Thoracic surgery. The given dataset is from 2007-2011. It consists of data of 470 lung cancer patients which were scaled on 17 variables to anticipate life after Thoracic surgeries within 1 year of surgery of people between ages of 21-87 years. WEKA toolkit (version 3.6.11) is used for data interpretation.

In the study various algorithms are tested out of which best possible algorithms are compared to get the best possible solutions to anticipate the Post- operative life of lung cancer patients after 1 year.Do to so only the. The Three Machine Learning Techniques were considered which were naive Bayes, logistic regression, and random forest so that the ML Model gives solutions on the basis of accuracy and feasibility.

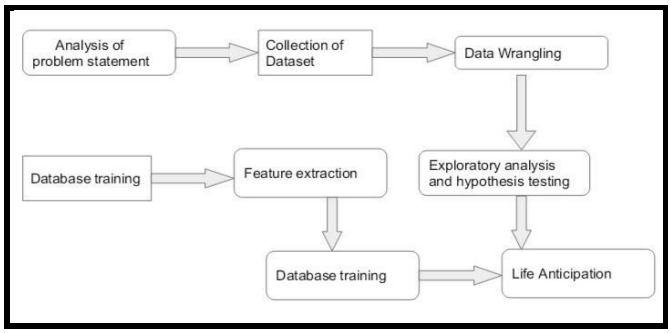

Fig 1. Flow chart depicting mapping of life anticipation of patients after thoracic surgery

Following steps were followed while developing the Machine Learning model -:
1. The current problem scenario is student
2. Various Functional requirements required for the Machine learning model were collected.
3. The machine learning model is evaluated and mapped on the basis of its feasibility..
4. Previous related works are analyzed.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

5. Different Machine learning techniques are selected to map and analyse on different parameters for building the model.
6. Selecting an algorithm development approach.
7. The data mining techniques which are selected are further compared on different scales and analyzed;
8. Essential Coding is performed to build the Ml Model by using the various data mining techniques which were selected.

1. Data Wrangling

The goal of Data Wrangling is to remove the unnecessary Id columns from the data set because it didn't contain the patient's description, the columns were converted so that it can be more human readable. It was observed that no missing values were there in the given dataset chosen. When data was analyzed 16 outliers were noticed. These 16 outliers were ignored and were removed after data wrangling because they were unnecessary for the data set; the instances were now reduced to 454 for the new dataset to be considered.

A. 2. Exploratory analysis of Data

The new dataset containing 454 instances was analyzed and it was seen that there was a total of 15.20% rate of death within 1 year after Thoracic surgery. Mean difference % between dead and live 1 year is plotted to determine which feature has to be focused. It was seen that Features like Dyspnoea,Pain,PAD, Diabetes mellitus had higher significance. Further for each attribute mean differences were compared for Diagnosis, performance and tumor size attribute. To determine and calculate the significance for the model more accurately hypothesis testing was done and its results were then seen.

A. 3. Hypothesis testing

It was seen that when Hypothesis testing was performed the level of significance calculated was 0.05 which depicted that attributes which have a significance lesser than 0.05 will be rejected and not considered. So the Null hypothesis revealed that for attributes FVC, FEV1, Pain, Haemoptysis, Weakness, MI_6mo, PAD, Smoking, Asthma and Age were not rejected in hypothesis testing because the level of significance was higher than 0.05.Pearson correlation was calculated for FVC and FEV1 which was 0.89 to merge two columns into one to minimize the effect of FVC and FEV1.

A. 3. Machine Learning techniques

Three data mining techniques were used: logistic regression, Random forest and Näve bayes to anticipate the life post thoracic surgery within 1 year. Which were evaluated on various basis accuracy, F measure,ROC curve, Gmean, TNR and TPR.

B. Automated disease Analyzer at initial phases

The statistics of different type of diseases
is from UCI Repository. It consists of data of 42 different types of diseases having scaled on 132 parameters to anticipate chronic illness in initial phases. The dataset used is broken into 2 parts to further perform the training and testing for building the ML model. Symptoms are mentioned in 132 columns while the 133th column is for diagnosis. The symptoms are scaled on various diseases to predict the disease.

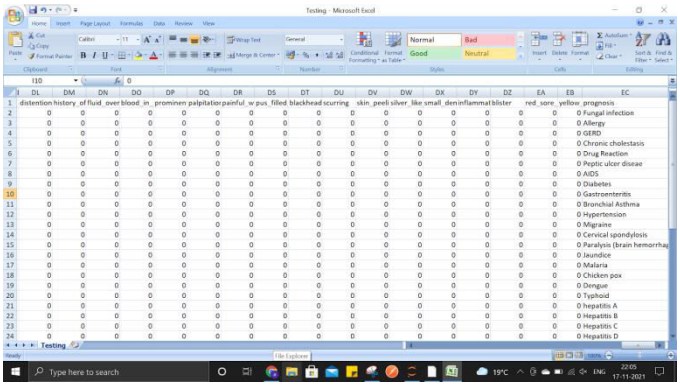

Fig 2. Column id showing description of patients

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

Three algorithms which are tested to analyse the ML model are decision trees, random forests and Näive bayes. The three algorithms are used to get similar predicted diseases which means 2 out of 3 algorithms should give the same predicted diseases when mapped on given symptoms by user input.

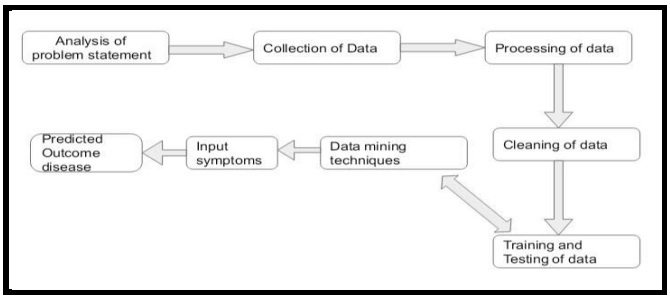

Fig 3. Flow chart depicting mapping of how diseases can be predicted using ML model.

Steps that were followed while developing and executing the Machine Learning model -:

1. Collection of Dataset
2. Analyzing the data set
3. Data Wrangling
4. Training and testing the data set
5. Training the model on different data mining techniques
6. Mapping the symptoms of different diseases.
7. Coding is performed to map the different ML techniques that were taken into consideration..
8. Providing input fields with the symptoms on the GUI which is built on Tkinter
9. Predicted disease can been seen on GUI built by providing the different inputs by the user.

.

B. 1. Data Cleaning

The goal of data cleaning is to rename Id columns with human readable names which are symptoms description. By doing this symptoms can be precisely seen with description.

B. 2. Training and testing the data

The data is trained and tested on various symptoms mapped for various diseases

B. 3. Data mining techniques

The symptoms are mapped on 42 different diseases by using decision Tree, random forest and Näve bayes algorithm.

B. 4. User input

Input fields are provided by the user on the GUI built and predicted disease is calculated by using the data mining techniques.

IV. EXPERIMENTAL RESULTS

A. Prediction of life post thoracic surgery within 1 year

Various Machine Learning techniques o were inspected on basis of accuracy, F measure,ROC curve, Gmean, TNR and TPR. ROC curve is used to calculate the built model Performance, whereas Gmean measures the machine learning model quality rate

The following mathematical expression is used to calculate the accuracy of each machine learning algorithm so that the best possible result can be used to anticipate post operative Life after Thoracic surgery as well as early detection of various diseases.

$$\text{Accuracy} = TP+TN/(P+ N). \quad (1)$$

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

TABLE II. THE FOLLOWING TABLE SUMMARIZES THE CONFUSION MATRIX

		Predicted Outcome
Actual value	P	TP , FN
	N	FP , TN

Accuracy may or may not give precise results while building ML models ; this often results in an imbalance of data sets that have been chosen. The UCI Repository which is the thoracic surgery dataset has 70 true instances and 400 false instances because it is an imbalance data. Because of this to measure accuracy F measure,TNR,TPR and ROC curve is used.
Mathematical expression that is taken into account to calculate F measure is shown below:
F Measure = 2TP/(2 TP+FP+FN) (2)

Further we also calculate Gmean to measure quality. Following mathematical expression is used:
Gmean = TPRTNR (3)

Further TPR and TNR values are also calculated:
TPR= TP/(TP+ FN) (4)

TNR= TN/(TN+ FP) (5)

The data set is splitted into 2 parts where 90% portion is for training the given dataset and other remaining for testing purpose. The classifiers are taken into consideration to anticipate the outcome. It is achieved by doing the training of data after this the classifier is used in testing.

TABLE III. THE FOLLOWING TABLE SHOWS THE PREDICTION MEASURES COMPARISON FOR DIFFERENT MACHINE LEARNING TECHNIQUES ON DIFFERENT PARAMETERS

ML Techniques	**Accuracy**	**F-measure**	**ROC curve**
Logistic Regression	80.91%	0.81	0.51
Random Forest	88.73%	0.834	0.683
Naive bayes	84.51%	0.829	0.738

By performing training as well as testing on the data by using various Machine learning algorithms the accuracy was mapped and calculated which gave the anticipated results. It was seen that the Random forest algorithm gave an accuracy of 88.73% , logistic regression gave an accuracy of 80.91% and Naive Bayes gave an accuracy of 84.51%. We use F measure and ROC curve because accuracy alone may not be sufficient as previously stated
ROC curve is simply taken into consideration to calculate the built model Performance which means that ML model is further analysed to see that whether using ROC curve will give better result or not.. Test accuracy is determined by F-measure.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

In ROC curve false positives division is mapped.In terms of talking about ROC curve naïve Bayes was better than Random forest and Logistic regression. Whereas in terms of accuracy if the data is not imbalanced Random forest algorithm was better than Logistic regression and naïve Bayes technique.

B. Automated disease Analyzer at initial phases

Machine Learning techniques were inspected on different parameters. To do so different diseases were mapped on different symptoms to predict the disease. The three machine learning algorithms used are Decision tree, Naive bayes,Random forest. It was seen that if only two algorithms were used the accuracy was 50-50 % it didn't give similar predicted disease if symptoms were varied. But when other algorithms were taken into consideration a similar pattern or similar results were seen for a disease. The GUI built is interactive with respect to the user as users have free liability to input the symptoms but however symptoms are chosen randomly the result may show diversion in the result.

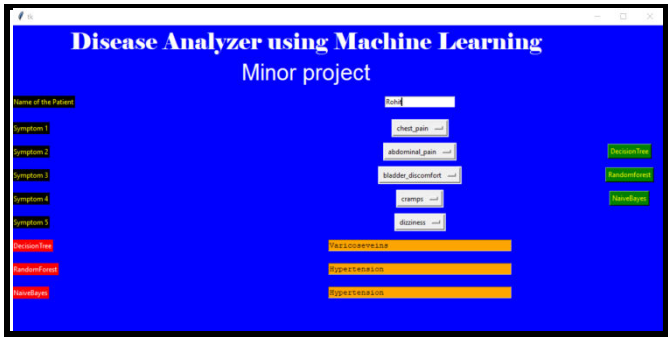

Fig 4, GUI of disease prediction ML model.

Figures shows the results of text detection from an image and inpainting by using exemplar based Inpainting algorithm. Figs. 2, 3, 4 (a) shows the original image. (b) is the image obtained by applying first set of criteria. All objects whose area greater than 10000 and filled area greater than 8000 are eliminated and major axis lengths are in between 20 to 3000 are considered to be text. Still, some small non-text objects are detected.

VI. FUTURE SCOPE

The results concluded at the end of the study reveals that there might be some restrictions also such as the data set taken from UCI Repository is from 2007-2011 which is the past dataset. If Recent dataset is used the results might alter depending upon various internal and external factors. Further it's seen that the data set might be restricted to one nation or area only.

V. CONCLUSION

The three machine learning algorithms are used to anticipate life post thoracic surgery as well as for early detection of various diseases at initial phases. The three data mining techniques were compared, mapped and scaled on different parameters; it was noted that accuracy might not give precise results every time if data is imbalanced. In this case ROC curve is taken into consideration to have a better understanding of which data mining technique should be used. For further studies and research purposes in future all machine learning algorithms should be taken into consideration to see what are their possibilities for predicting diseases.

REFERENCES

[1] V. Sindhu, S. A. S. Prabha, S. Veni , and M. Hemalatha, “Thoracic surgery analysis using data mining techniques” , International Journal of Computer Technology & Applications , vol. 5, pp 578-586, May, 2014.

[2] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, “Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal, vol 13, pp 8-17, 2015.

[3] Adam Abdulhamid, Ivaylo Bahtchevanov, Peng Jia,“Life Expectancy Post Thoracic Surgery” ,Stanford University - CS 229,2014.

[4] Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma , “Machine Learning For Prognosis of Life Expectancy and Diseases” ,International Journal of Innovative Technology and Exploring Engineering ,Vol. 8, August 2019.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012083|

[5] Kittipat Sriwong, Kittisak Kerdprasop, and Nittaya Kerdprasop,"Post-Operative Life Expectancy of Lung Cancer Patients Predicted by Bayesian Network Model" , International Journal of Machine Learning and Computing, Vol. 8, June 2018.

[6] Lavesh S, P Sree Lekha, Naveen Kumar R, Manoj T V, Sushila Shidnal , "Lung Cancer Detection and Life Expectancy Post Thoracic Surgery Using CNN and Supervised Machine Learning Algorithms" , International Research Journal of Engineering and Technology , Vol.07 ,Apr 2020.

[7] Abeer S. Desuky and Lamiaa M. El Bakrawy ,"Improved Prediction of Post-operative Life Expectancy after Thoracic Surgery" , Advances in Systems Science and Application, Vol.16, Jan 2016.

[8] M. Zięba, J .M. Tomczak, M. Lubicz, and J. Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients" , Applied Soft Computing, vol. 14, pp 99-108, Jan. 2014.

[9] Md. Ahasan Uddin Harun and Md. Nure Alam, "Predicting Outcome of Thoracic Surgery by Data Mining Techniques", IJARCSSE, vol. 5, no. 1, pp 7-10, 2015.

[10] Mark A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning", International Conference on Machine Learning, pp 359-366, 2000.

[11] Vishal Gupta, et al., "Performance of Various Feature Selection Techniques under Loaded Networks", International Journal of Computer Applications, vol. 78, no. 2, September 2013.

[12] S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and na¨ıve bayes algorithms, "International Journal of Science, Engineering and Technology Research" Vol. 4, April, 2015.

[13] Y. Khourdifi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, "Int. J. Intell. Eng. Syst", Vol. 12, Oct, 2019.

[14] S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, "International journal of environmental research and public health" , Vol.15, Jul, 2018.

[15] S. Mohan, C. Thirumalai, G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" , IEEE Access , Vol. Jun (2019).

[16] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar,T.P.R. Vital, "Intelligent parkinson disease prediction using machine learning algorithms", International Journal of Engineering and Innovative Technology, Vol. 3, Sep, 2013.

[17] Kumar, C. Sunil and R. J. Sree. , "Application of ranking based attribute selection filters to perform automated evaluation of descriptive answers through sequential minimal optimization models", Ictact Journal on Soft Computing: Special Issue on Distributed Intelligent Systems and Spplications, Vol. 05, Oct, 2014.

[18] Jasmina Novakovi´c, Perica Strbac and Dusan Bulatovi´c., " Toward optimal feature selection using ranking methods and classification algorithms", Yugoslav Journal of Operations Research, Vol. 21, No. 1, pp.119-135, Mar, 2011.

[19] R.D.H.D.P. Sreevalli, K.P.M. Asia, "Prediction of diseases using random forest classification algorithm", Zeichen Journal, Vol. 6 , Apr 2020.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details