International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

The Real Time Lung Cancer Stages Prediction using KNN and Decision Tree

Leela L O[1] and Mr Manjunatha H T [2]
Student, Department of Computer Application, Jawaharlal Nehru College of Engineering, Shivamogga[1].
Assocoate prof, Department of Computer Application, Jawaharlal Nehru College of Engineering, Shivamogga[2].

ABSTRACT

To identify or forecast the stage of lung cancer, researchers are utilizing machine learning to predict lung cancer. Due to their inability to understand the ailments they are afflicted with, many people in our day to day life suffer from a variety of illnesses. The most widespread diseases in the world is lung cancer. This is one of the second most common cancers to be diagnosed. In an effort to solve this issue, this work was created. In order to save lives, early cancer detection is crucial. Building a tool for lung cancer screening using Python machine learning is the project's major goal. A patient's chances of healing and recovering are improved with early detection. Effective cancer detection in the body is greatly aided by technology. This work is used to predict the stages of Lung Cancer. Doctor can send the report of the Patient.

Keywords: Lung Cancer, python, Machine Learning, KNN Algorithm, Decision Tree algorithm.

1. INTRODUCTION

A category of illnesses known as cancer is redefined by the unchecked growth and division of aberrant cells. People may die if the spreading is not stopped. The most prevalent disease in the world, lung cancer accounted for a significant portion of all newly diagnosed cases in 2018. Early cancer diagnosis in people is possible with the use of the machine learning algorithm for lung cancer detection. This can be highly beneficial for physicians and radiologists in providing patients who contact them with better results. These methods have the potential to cure people, restrict the spread of lung cancer, and save millions of lives. More such advancements in healthcare can be made using machine learning algorithms.

One of the most prevalent diseases in the world is lung cancer. Although lung cancer cannot be averted, the risk can be decreased. Thus, a patient's chance of survival depends on the early identification of lung cancer. Lung cancer incidence is directly inversely correlated with the frequency of heavy smokers. This work uses Machine-learning algorithm to identify lung cancer in this research utilizing datasets on the symptoms. Python is one of the effective programming languages used in this work. This initiative can be utilized to treat those who have lung cancer and have been diagnosed early and assist them in overcoming this illness.

According to statistics in 2018, it was predicted that lung cancer caused 9.6 million deaths. As early a diagnosis of cancer as possible is crucial since it tends to spread and, in the event of a bigger spread, is incurable. Since symptoms only appear in the latter stages of lung cancer, it is challenging to get a diagnosis. It is also practically hard to save a person's life in this stage. Lung cancer symptoms, which can include a

chronic cough, bloody sputum, chest pain, voice changes, and increased shortness of breath, typically do not manifest until the disease has progressed. The single biggest risk factor for lung cancer is smoking cigarettes. Smoking still accounts for 90% of lung cancer fatalities. In order to classify the cancer stages like low, medium and high along with report using KNN and Decision tree algorithm.

Classifier uses the features like Genetic_Risk, Occupational_hazard, chest_p ain, chronic_lung_cancer,clubbing_of_fingure_nail,coughing_of_blood,dry_cough,fatigue, passive_smoker,smocking,weight_loss. Based on these features it effectively predicts the stages of lung cancer.

2. LITERATURE SURVEY

Anand, A. and Shakti, D et al [1], this paper defines the Healthcare has always been having been a sensitive subject for all of us. We can live better lives if we can predict numerous health conditions in advance. Numerous conditions might are numerous conditions that might cause health issues, including cancer, heart disease, diabetes, arthritis, pneumonia, lung illness, liver disease, and brain disease, all of which have substantial risks. Appropriate predictive models that are appropriate are essential to lower the risk of health problems. Thus, it became a motivating element for the authors to extensively review the body of research on the subject and, as a result, to discover appropriate machine learning approaches so that advancement may be achieved when choosing a prediction model. The idea of a survey is used in this chapter to present prediction models for healthcare-related problems as well as the difficulties with each model. The following will be extensively covered in this chapter: Health industry applications of machine learning include research on numerous cancers, heart, brain, and other illness predictions models as well as a comparison of different machine learning prediction methods.

Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H et al [2], this paper defines the application of information technology is quite difficult to comprehend. It plays a part in a variety of fields, along with the foretelling of diseases like diabetes. Today, diabetes is extremely harmful all around the world. Information technology can play a significant role in offering treatment to reduce the numbers who pass away from such a disease. Technology is powerful in various ways across various industries. most of the time, solving the issue means that it is half solved. In this study, four algorithms are presented: ZeroR, SVM, J48, and NB.

Kanchan, B.D. and Kishor, M.M et al [3], this paper defines the Lung cancer is the leading cause of death, according to a global assessment of the disease's causes of death. Unfortunately, the healthcare sector gathers a variety of material about heart disease that is not "mined" to find hidden information for smart decision-making. This research of PCA determines the bare minimum of attributes necessary to improve the precision of various supervised machine learning methods. The goal of this study will be to investigate supervised machine learning techniques for lung cancer disease prediction. Conflation and preparation are two crucial strategies used in data mining. Diabetes is a serious illness that can be prevented in many developed and emerging nations, including India.

Kazeminejad, A., Golbabaei, S. and Soltanian-Zadeh, H et al [4], this paper defines the Lifestyle diseases are conditions that are linked to a person's or a population's way of life. Unfortunately, the vast amounts of disease-related data collected by the healthcare sector are not mined for any hidden insights that can help with decision-making. Long-term activity monitoring would aid in the management of diseases linked to a sedentary lifestyle. These illnesses are frequently connected to a person's manner of life. The likelihood of developing such diseases in later life is influenced by an unhealthy and erratic quality of living. People with erratic lives frequently experience the symptoms and early warning indications of these diseases. Using long-term activity monitoring, we suggest a unique healthcare strategy in this research to manage lifestyle diseases.

Milgram, J., Cheriet, M. and Sabourin, R et al [5], this paper offer a dual-process model of taxonomic and thematic similarity assessment that may be applied in machine learning applications based on recent research in the area of human similarity perception. In contrast to theme reasoning, which is primarily connected to metric distances, taxonomic reasoning is connected to predicate-based measures. We propose a method that fuses the two processes into a single kernel with similar characteristics. Using a greedy algorithm, an ideal measure is chosen for each feature dimension of the observational data: The measure that results in increased classification performance across the board for the entire model is identified after a variety of potential measures have been tried. These measurements are merged via generalization and quantization into a single SVM kernel.

Mishra, A.K., Keserwani, P.K., Samaddar, S.G., Lamichaney, H.B. and Mishra, A.K et al [6],this paper shows the lifestyle diseases are diseases that are influenced by a person's or a group's way of living. Unfortunately, the vast amounts of disease-related data that the healthcare sector accumulates aren't mined for any hidden insights that would aid in making wise choices. Long-term activity monitoring of a person would aid in the management of diseases linked to a sedentary lifestyle. These illnesses frequently have a link to a person's way of life. The likelihood of developing these diseases in later life is influenced by an unhealthy and erratic quality of living. People with irregular lifestyles frequently exhibit the earliest symptoms and indicators of these disorders.

Pattekari, S.A. and Parveen, A et al [7], this paper contains the very vast amount of data must be investigated in order to obtain any meaningful information, which is known as data mining. To study the many types of disease, various data mining approaches are used, including association rule mining, classification, and clustering. Data mining must deal with the issue of classification. A classifier performs a succinct and precise definition for each class that can be used to categorize succeeding records given that a database has a collection of records with a single class label on each one. One of the most hazardous diseases in the world is lung cancer. Lung cancer can be entirely cured if it is discovered early. By using Artificial Neural Networks and Naive Bayes to a sizable amount of healthcare data, data mining performs a useful role.

Sayali Ambekar and Dr.Rashmi Phalnikar et al [8], this research makes an effort to assist someone in predicting the condition they are experiencing based on their symptoms and the necessary readings of their bodily vitals. There are occasions when people continue to ignore health issues because of expensive medical costs. Later, serious problems from this could even result in death. Medical expenses might be a threat if they

are not covered by insurance. This website uses disease severity and type estimation to lessen the effort required of a typical person. Using various machine learning methods, this article developed a disease prediction system. The diagnosis system generates an output indicating whether or not the user has the specific condition based on the user's symptoms, age, and gender. Exercise regimens and food plans that can at least somewhat lessen the consequences of the disease are also offered, depending on its severity. It offers a straightforward yet efficient method for anticipating the sickness if the provided

Sharma, M. and Mujumdar, P.K et al [9], this paper contains the term "lifestyle diseases" refers to conditions whose development is mostly based on a person's daily routine and results from an unsuitable interaction between that person and their environment. Bad eating habits, physical inactivity, poor body mechanics, and biological clock disturbances are the main causes of lifestyle disorders. India will lose an estimated $236.6 billion by 2015 as a result of unhealthful habits and poor eating, according to a report co-authored by the World Health Organization (WHO) and the World Economic Forum.

Suzuki, A., Lindor, K., St Saver, J., Lymp, J., Mendes, F., Muto, A., Okada, T. and Angulo, P et al [10], this paper includes the uncertainty surrounds the effects of lifestyle changes on nonalcoholic fatty liver disease. We sought to ascertain if variations in lifestyle and body weight were related to variations in serum ALT levels. Data from 1546 employees' annual health examinations are analyzed in this research. By removing those who had other liver diseases, we chose 348 male subjects out of the 469 subjects with increased ALT. They were observed for a year to evaluate the relationship between changes in lifestyle and changes in serum ALT. To determine the relationship between lifestyle management and persistently normal ALT, the 136 subjects whose ALT had normalized were followed for two years.

3. METHODOLOGY

The Proposed methodology contains following stages:

i. **Disease:** It is a dataset which includes number of records as the patients symptoms
ii. **Data Processing:** Data processing must be done appropriately in order to avoid having an adverse effect on the final product or data output after data is collected and converted into usable information.
iii. **Training set:** An initial dataset known as a "training dataset" is used to train machine learning models to recognize specific patterns or carry out specific tasks.
iv. **Testing set:** An additional data set used to test a machine learning algorithm after it has been trained on an initial training data set is known as a test set in machine learning.
v. **Feature-selection:**Genetic_Risk,Occupational_hazard,chest_pain,Chronic_lung_cancer, clubbing_of_fingure_nail,coughing_of_blood,Dry_cough,fatigue,passive_smoker, smocking, weight_loss these are the feature selected by doctor..
vi. **Ensemble classification Technique:** It is committee-based learning, or the training of numerous classifier systems that use multiple hypotheses to address the same issue.

vii.KNN: The KNN algorithm simply stores the dataset during the training phase and then classifies incoming data into a category that is very close to the previously stored dataset.

Step 1: Decide on the neighbours' K-numbers.

Step 2: Calculate the Euclidean distance between K neighbours in step two.

Step 3: Based on the determined Euclidean distance, select the K closest neighbours.

Step 4: Count the number of data points in each category among these k neighbours.

Step 5: Assign the fresh data points to the category where the neighbour count is highest.

Step 6: Our model is complete.

viii.Decision Tree: A supervised training method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a tree-structured classifier, where internal nodes stand-in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

Step 1: According to S, start the tree at the root node, which has the entire dataset.

Step 2: Utilize the Attribute Selection Measure to identify the dataset's top attribute

Step 3: Subset the S to include potential values for the best qualities.

Step 4: Create the decision tree node that has the best attribute in step four.Use the selections of the dataset generated in step 3 to iteratively develop new decision trees in **Step 5:** Continue along this path until you reach a point when you can no longer categorise the nodes and you refer to the last node as a leaf node.

ix.Dieses prediction: Finally using the dataset the stage of lung cancer is predicted based on symptoms.

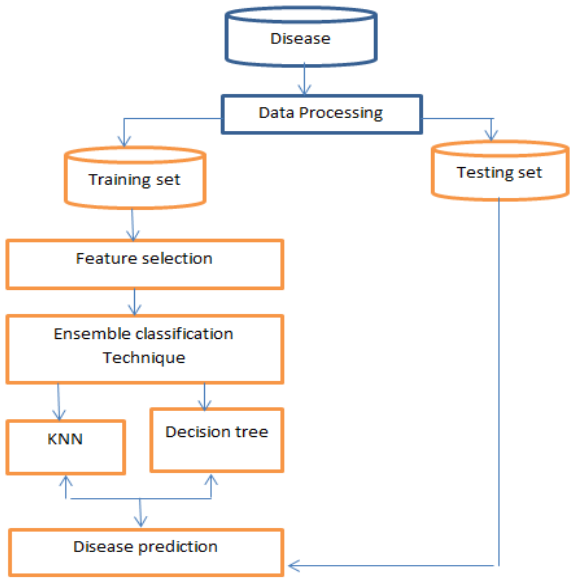

Fig: Block diagram of lung cancer prediction

Results

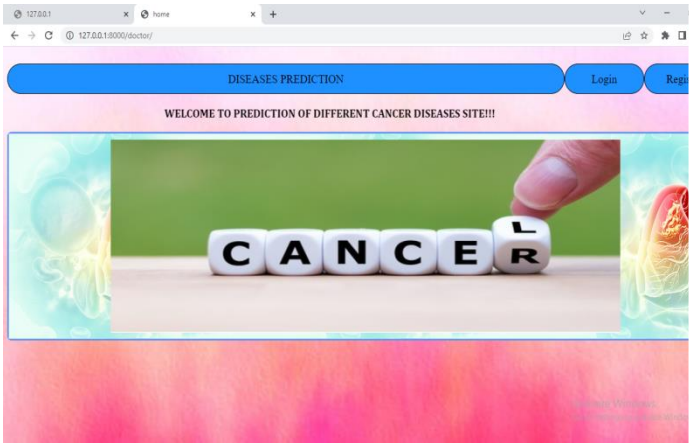

Fig1: This picture is refers to the Login Page of Doctor Module.

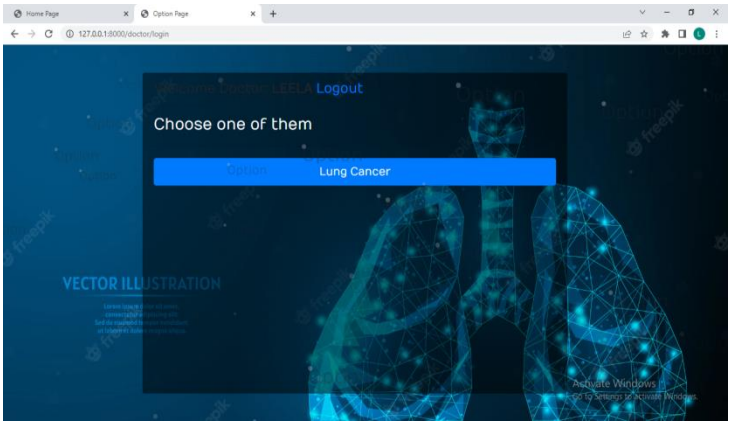

Fig 2: This picture is refers to the Home page of Doctor Module.

Fig 3: This picture refers to the Login Page of the Patient Module.

4. CONCLUSION

The primary objective of this project is to forecast the lung cancer stage. This initiative was created primarily for those who have lung cancer but are unaware that they are the only ones who have it. This project was created for such situations utilizing the machine learning KNN algorithm and Decision tree. People's inactivity, bad eating patterns, and stress at work are the main causes of most diseases. If we can find the causes of the diseases early on, we will be able to prevent them and this methodology will assist to prolong the lives of the patients. In this review paper, we seek to list all the significant studies that have been conducted in the recent years and can be enhanced to produce better outcomes.

REFERENCES:

[1] Anand, A. and Shakti, D., 2015. Prediction of lung cancer based on personal lifestyle indicators. In Next generation computing technologies (NGCT), 2015 1st international conference on (pp. 673–676). IEEE.
[2] Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among MiddleAged People Using Data Mining Techniques. Procedia Computer Science, 132, pp. 1068–1076.
[3] Kanchan, B.D. and Kishor, M.M., 2016. Study of machine learning algorithms for special disease prediction using principal of component analysis. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on (pp. 5–10). IEEE.
[4] Kazeminejad, A., Golbabaei, S. and Soltanian-Zadeh, H., 2017. Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI. In Artificial Intelligence and Signal Processing Conference (AISP), (pp. 134–139). IEEE.

[5] Milgram, J., Cheriet, M. and Sabourin, R., 2006. "One against one" or "one against all": Which one is better for handwriting recognition with SVMs?. Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule (France), Suvisoft, 2006.
[6] Mishra, A.K., Keserwani, P.K., Samaddar, S.G., Lamichaney, H.B. and Mishra, A.K., 2018. A decision support system in healthcare prediction. In Advanced Computational and Communication Paradigms(pp.156–167).Springer,Singapore.
[7] Pattekari, S.A. and Parveen, A., 2012. Prediction system for lung cancer disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp. 290–294.
[8] Sayali Ambekar and Dr.Rashmi Phalnikar, 2018. Disease prediction by using machine learning, International Journal of Computer Engineering and Applications, vol. 12, pp. 1–6.
[9] Sharma, M. and Mujumdar, P.K., 2009. Occupational lifestyle diseases: An emerging issue. Indian Journal of Occupational and Environmental medicine, 13(3), pp. 109–112.
[10] Suzuki, A., Lindor, K., St Saver, J., Lymp, J., Mendes, F., Muto, A., Okada, T. and Angulo, P., 2005. Effect of changes on body weight and lifestyle in nonalcoholic fatty lung cancer disease. Journal of Hepatology, 43(6), pp. 1060–1066.
[11] Manjunatha HT and AjitDanti. "A Novel Approach for Detection and Recognition of Traffic Signs for Automatic Driver Assistance System Under Cluttered Background" - Recent Trends on Image Processing and Pattern Recognition, Springer Nature Singapore, Pte Ltd. 2019, RTIP2R 2018, CCIS 1035, pp. 1–8, 2019, ISBN 978-981-13-9181-1 DOI -https://link.springer.com/chapter/10.1007/978-981-13-9181-1_36.
[12] Manjunatha HT and Ajit Danti. "Detection and Classification of Potholes in Indian Roads using Wavelet Based Energy Modules"IEEE- 978-1-5386-9319-3/19 © 2019 ,SCOUPS Nature .
[13] Manjunatha HT, Ajit Danti, Arunkumar KL, Rohith D" Indian Road Lanes Detection Based on Regression and clustering using Video processing Techniques", 3rd International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R,2020). Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Springer, Scopus Indexed 3rd and 4th January 2020. Springer Nature Singapore Pte Ltd. 2021 K. C. Santosh and B. Gawali (Eds.): RTIP2R 2020, CCIS 1380, pp. 1–14, 2021. https://doi.org/10.1007/978-981-16-0507-9_17
[14] Manjunatha HT and AjitDanti," Indian traffic sign board recognition using Normalized Correlation Method", International Journal of Computer Engineering and Applications (IJCEA), Volume XII, Issue III, March 18, ISSN 2321-3469
[15] Manjunatha HT and AjitDanti, "Segmentation of Traffic Sign Board in a cluttered background using Using Digital Image Processing", National Conference on Network Security, Image Processing and Information Technology, March 2017.
[16] K L Arunkumar and Ajit Danti. H T Manjunatha, D Rohith "Classification of Vehicle Type on Indian Road Scene Based on Deep Learning": Recent Trends on Image Processing and Pattern Recognition, Springer Nature Singapore Pte Ltd. 2021, RTIP2R 2020, CCIS 1380, Springer, pp. 1–10, 2021.
[17] Arunkumar K L, Ajit Danti, Manjunatha H, "Classification of Vehicle Make Based on Geometric Features and Appearance-Based Attributes Under Complex Background" , Springer 1035 (CCIS), pp 41-48
[18] K L Arunkumar, Ajit Danti, "A NOVEL APPROACH FOR VEHICLE RECOGNITION BASED ON THE TAIL LIGHTS GEOMETRICAL FEATURES IN THE NIGHT VISION", International Journal of Computer Engineering and Applications, Volume XI

[19] Manjunatha HT, Arunkumar K L, Ajit Danti, "A Novel Approach for Detection and Recognition of Traffic Signs for Automatic Driver Assistance System Under Cluttered Background", Springer 1035 (CCIS), pp 407-419
[20] KL Arunkumar, A Danti, HT Manjunatha, "Estimation of vehicle distance based on feature points using monocular vision", IEEE 8816996 (2019), 1-5
[21] KL Arunkumar, A Danti, HT Manjunatha, D Rohith, "Classification of Vehicle Type on Indian Road Scene Based on Deep Learning", Springer, Singapore 1380 (2021), 1-10
[22] Indian Road Lanes Detection Based on Regression and clustering using Video Processing Techniques, HT Manjunatha, A Danti, KL ArunKumar, D Rohith Springer, Singapore 1380 (CCIS), 193-206
[23] Recognition of Vehicle using geometrical features of a tail light in the night vision,
Arunkumar K L, Ajit Danti, National Conference on Computation Science and Soft Computing (NCCSSC-2018)