IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203064 |

Artificial Intelligence Framework for Early Diagnosis and Prediction of Lung Cancer

Cheedella Hemanth Satya Sai, Jagadeesh N, Kodati Tagur Vara Prasad, Vemula Sai Teja,

Ms.Arockia Rosy, B.Tech,M.Tech,

UG Students, Dept. of Information Technology, R.M.D. Engineering College, Tiruvallur, Tamil Nadu, India

Assistant Professor, Dept. of Information Technology, R.M.D. Engineering College, Tiruvallur, Tamil Nadu, India

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.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 to detect the lung cancer,The proposed model works with the efficiency of 95% accuracy.The proposed system is for predicting the chances of lung cancer by displaying namely yes or no Thus, mortality rates can be reduced significantly.

KEYWORDS: Artificial intelligence, Cancer prediction, Lung cancer, Random forest algorithm, disease, detection.

I. INTRODUCTION

The lungs are the main organs of respiration. The human body has two lungs, one on each side of the chest. The left lung is smaller than the right, leaving room for the heart. During breathing, the chest rises and falls. That is because by inhalation, the lungs swell, and by exhalation, they shrink. The lungs are responsible for enriching the blood with oxygen. The heart sends to the lungs blood tha is low in oxygen and rich in carbon dioxide. The blood inside the lungs is "cleansed", absorbs oxygen and leaves carbon dioxide. Carbon dioxide is eliminated during exhalation, while oxygen enters the lungs during inhalation.

The use of machine learning algorithms can help healthcare professionals to identify patterns and make informed decisions quickly. In this project, we propose an AI framework for early diagnosis and prediction of lung cancer using the Random Forest algorithm.

The Random Forest algorithm is a machine learning model that falls under the category of ensemble learning methods. Ensemble learning involves combining multiple models to improve the accuracy and robustness of predictions. The Random Forest model combines multiple decision trees to make more accurate and stable predictions. Each decision tree in the forest makes an independent prediction, and the final prediction is based on the majority vote of all the trees. Random Forest algorithm to provide timely and accurate diagnosis of lung cancer. The proposed system will be a valuable tool for healthcare professionals to identify patients at risk of developing lung cancer, allowing for early intervention and improved patient outcomes.Cancer is becoming a common cause of death.

The algorithms such as logistic regression, decision trees, and support vector machines is used to train models on labeled data to classify patients as either having lung cancer or not. The framework utilizes a machine learning algorithm called Random Forest to analyze patient data and make predictions about their risk of developing lung cancer. Random Forest is a powerful algorithm that has been used successfully in several medical domains. It works by building multiple decision trees and combining their predictions to obtain a more accurate result. The project domain includes data collection, processing, and analysis. The framework will be trained on a large dataset of medical images, patient history, and demographic information to learn the patterns and characteristics of lung cancer. The dataset will be curated from various sources, including hospitals, clinics, and research institutions, to ensure that it represents a diverse population.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203064 |

II. RELATED WORK

RashmeeKohad, et al. [2018] 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 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**, [2018] 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%.

Supreet Kaur, et al. [2019] 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, 5 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.

JaneeAlam, et al. [2020] 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%.

III. PROPOSED METHODOLOGY

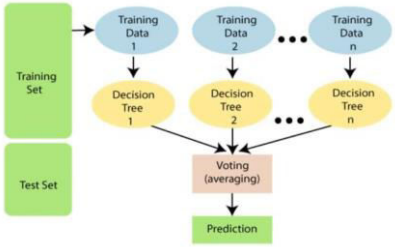
The proposed methodology is based on a dataset consisting of features that capture human habits (such as smoking, and alcohol consumption) and signs/symptoms as risk factors that lung cancer patients usually incur.However, these signs are not necessarily related to lung cancer disease.Unlike other cancers, lung cancer cannot be seen with the naked eye, and its symptoms are often accompanied by other disease symptoms. The most frequent symptoms are allergies, asthma, shortness of breath, and coughing. The Random Forest Algorithm is a machine learning algorithm that is particularly well- suited for the analysis of large datasets. It works by creating multiple decision trees, each of which analyzes different subsets of the data. The algorithm then combines the results of these decision trees to make a final prediction. This approach is particularly effective at identifying complex relationships between different features of the data, which is important in the diagnosis of lung cancer.

The proposed AI framework utilizes various machine learning algorithms and techniques, including feature selection, and dimensionality reduction, to process and analyze medical imaging data, clinical information, and genetic data for lung cancer diagnosis and prediction The proposed AI framework has the potential to revolutionize lung cancer diagnosis and prediction, leading to improved patient outcomes and reduced healthcare costs.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203064 |



Advantages:

1. Identify early signs of lung cancer:By analyzing large amounts of patient data, including demographic information, medical history, and diagnostic test results, the algorithm can identify patterns and relationships that may be indicative of early-stage lung cancer. This can lead to earlier and more accurate diagnosis, which can improve patient outcomes and reduce healthcare costs.

2. Ability to learn and adapt:As new patient data becomes available, the algorithm can incorporate this data into its analysis, which can improve the accuracy and reliability of its predictions This adaptability is particularly important in the field of lung cancer diagnosis, where new diagnostic techniques and treatments are constantly being developed.

3. Accuracy:Accuracy is high and we can know the cancer risk at low cost.

Dataset Splitting :

The dataset is divided into two subsets a training set and a test set. A common practice is to split the dataset into 80% for training and 20% for testing. This division ensures that the model is trained on a substantial portion of the data while reserving a separate portion for evaluating its performance.

Pandas :

Pandas is a popular Python library for data manipulation and analysis. It provides data structures like Series and DataFrame for handling structured data, supports data input/output from various formats, and offers powerful tools for data cleaning, transformation, and analysis. NumPy : NumPy is often used in image processing tasks, where images are represented as arrays of pixel values. It enables various operations, like filtering, transformations, and manipulations of image data.

Matplotlib :

Matplotlib is a popular and widely-used data visualization library in Python. It provides a flexible and comprehensive set of tools for creating a wide range of static, animated, and interactive plots and charts for data visualization.

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

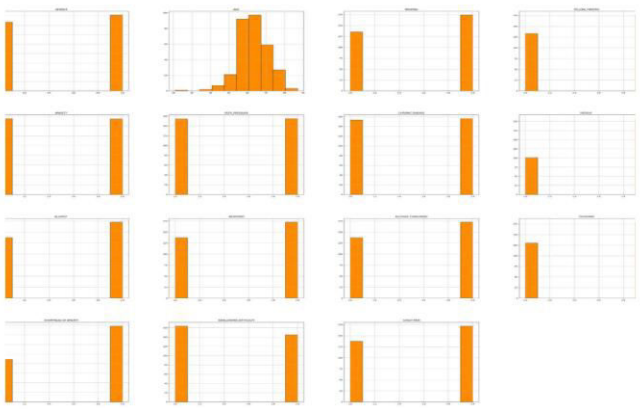
TP - True Positive , TN - True Negative, FP - False Positive

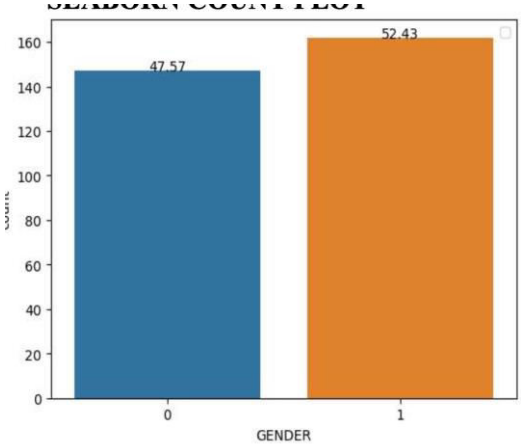
IV. RESULTS

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203064 |





The proposed system is more efficient and gives the more appropriate results compared to the existing systems. In the proposed system we are using the Random Forest algorithm that combines multiple decision trees to improve the accuracy of the model. One of the key advantages of random forest algorithm is its ability to handle large and complex datasets with high dimensionality. The model is created following 4 steps;Data Collection,Preprocessing and feature extraction,Training and validation and Testing.First the dataset has import and also import libraries that are needed for implementation of code we are using sigmoid activation function in the output layer because the output needs to be a probability value between 0 to 1.Binary accuracy as evaluation metric.Then the model is completed.accuracy thus obtained is 95.1%.

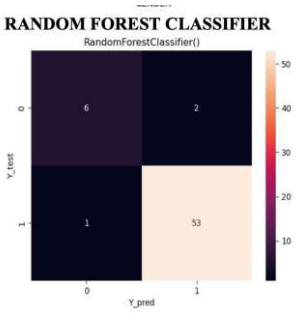


Fig no.1 Random Forest Classifier

Fig no.2 Accuracy Score

V. CONCLUSION AND FUTURE WORK

The development of an framework for early diagnosis and prediction of lung cancer using the random forest algorithm is a promising area of research. The use of machine learning algorithms such as random forest has the potential to significantly improve the accuracy and speed of lung cancer diagnosis, which can lead to better patient outcomes and improved healthcare efficiency.The results of this project demonstrate that the random forest algorithm can effectively identify patterns and biomarkers that are associated with lung cancer, which can be used to develop predictive models for early detection and diagnosis. The algorithm's ability to handle large and complex medical datasets, high accuracy, make it an efficient and valuable tool for predicting lung cancer.Accordingly,the results proved that the system was efficient in providing a better accuracy as 95.1%.

In future, There are several potential enhancements that could be made to the artificial intelligence framework for early diagnosis and prediction of lung cancer. One possible improvement could be to incorporate additional data sources, such as genetic information or environmental factors, to improve the accuracy of the model. Another approach could be to explore more advanced machine learning techniques, such as deep learning, to improve the sensitivity and specificity of the diagnostic and predictive algorithms. Additionally, ongoing data collection and analysis could be used to refine and optimize the framework over time, leading to improved outcomes for patients and healthcare providers alike.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203064 |

In the future, advancements in the artificial intelligence framework for early diagnosis and prediction of lung cancer hold immense potential. Incorporating additional data sources beyond traditional patient datasets, such as genetic information or environmental factors, could enrich the model's predictive capabilities. Genetic data could unveil predispositions and molecular markers, while environmental data might highlight exposure risks.

REFERENCES

1. 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 2018.
2. 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 2018.
3. Supreet Kaur and Amanjot Kaur Grewal, "A Review Paper on Data dsMining Classification Techniques For Detection Of Lung Cancer", International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 11 Nov - 2019.
4. 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 2019.
5. 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 2019
6. 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) 2020.
7. Janee Alam1, Sabrina Alam and Alamgir Hossan, "Multi -Stage Lung Cancer Detection and Prediction Using Multi -class SVM Classifier" on 12 March 2020.
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, 2021.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details