IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.115112

LUNG CANCER PREDICTION USING MACHINE LEARNING

VIJAYARAGAVAN K [1], Dr. A.R. JAYASUDHA [2]

II MCA Student - MCA, Hindusthan College of Engineering and Technology, Coimbatore, India. [1]

Professor & Head-MCA, Hindusthan College of Engineering and Technology, Coimbatore, India.[2]

Abstract: Lung cancer is a kind of cancer that originates in the lungs and cannot be prevented in its late stages of development, but its risk can be reduced using the sessions. As a result, quick detection of lung cancer can help to lower the survival rate. The number of chain smokers is probably equal to the number of people affected by lung cancer. The lung cancer is predicted using the Logistic Regression. The study employs logistic regression to analyse categorical datasets. After evaluating parameters and assessing the significance of each influencing attribute, the model undergoes testing. This process yields 18 prediction models and identifies factors correlating with disease size risk. Utilizing logistic regression, the study predicts lung cancer occurrence in patients based on various factors such as symptoms, habits, and health history. Notable symptoms associated with lung cancer include smoking, alcohol consumption, swallowing difficulties, coughing, chronic ailments, fatigue, and age.

Keywords: Prediction, Logistic Regression, Machine Learning.

I. INTRODUCTION

1.1 PROBLEM DESCRIPTION

Lung cancer is a prevalent and life-threatening disease that originates in the lungs, often with limited opportunities for prevention in its advanced stages. However, early detection holds promise for improving survival rates. This study aims to explore the potential of logistic regression, a predictive modelling technique, in identifying individuals at risk of developing lung cancer based on a range of factors including symptoms, habits, and medical history. The research leverages logistic regression, specifically tailored for categorical datasets, to build predictive models. Through rigorous parameter assessment and significance testing of influencing attributes, the study seeks to construct a robust model capable of accurately predicting the likelihood of lung cancer occurrence. By examining correlations between disease size and various risk factors, including smoking, alcohol consumption, swallowing difficulties, coughing, chronic illnesses, fatigue, and age, the research endeavours to uncover insights into the complex interplay of these variables in lung cancer development. The ultimate aim is to create a predictive framework accessible to healthcare providers. This framework enables the assessment of individual risk profiles, allowing proactive intervention to stall the advancement of lung cancer. Early identification of high-risk individuals promises to enhance patient outcomes and alleviate the impact of this formidable illness.

1.2 OBJECTIVE

Creating a predictive model that can accurately identify and categorize possible cases of lung cancer based on pertinent data is the goal of a mini-project on lung cancer prediction. Machine learning algorithms and data analytics approaches are frequently employed to examine various aspects such as patient demographics, medical history, and diagnostic test results. The ultimate objective is to develop a trustworthy tool that will help medical professionals identify lung cancer early, enable timely intervention, and enhance patient outcomes overall. By offering a predictive tool to assist in the prompt identification of lung cancer, This mini-project aims to propel medical research and healthcare forward, laying the groundwork for streamlined and targeted treatment methods.

1.3 SCOPE

A mini-project focused on predicting lung cancer risk holds promise for significant progress in early detection and intervention strategies. Such a study seeks to create predictive models that can recognize trends and risk factors linked to lung cancer by utilizing machine learning algorithms and examining pertinent statistics. To improve prediction accuracy, the scope includes examining a variety of data sources, such as imaging and medical records. If this mini-project is completed successfully, this could lead to the development of a valuable tool tailored for medical professionals, facilitating swift and personalized interventions for individuals at risk of lung cancer. Ultimately, such advancements promise to enhance patient outcomes and alleviate the weight of this life-threatening illness.

II. LITERATURE SURVEY

• Sharmila Nageswaran, 1 G. Arunkumar, 2 Anil Kumar Bisht, 3 Shivlal Mewada, 4 J. N. V. R. Swarup Kumar, 5 Malik Jawarneh, 6 and Evans Asenso, "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing." 7. provides a thorough examination of how machine learning and image processing methods are applied to the identification and prognosis of lung cancer. The literature review champions state-of-the-art technologies such as CAD systems, emphasizing the critical role of early detection in combating lung cancer. It looks at image processing techniques for evaluating CT scans of the chest, emphasising segmentation and noise reduction to enhance image quality. The paper also looks at the use of machine learning, specifically ANN and K-means clustering, for the prediction and classification of lung cancer. Restrictions including the requirement for sizable datasets and difficulties with generalisation are also covered. Overall, the review sets the context for the research in the paper by offering a comprehensive overview of existing methods for machine learning and image processing-based lung cancer detection and prediction.

• “Kernel-based learning and feature selection analysis for cancer diagnosis” by Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkadar Benyettou, In his research, Mohammed Ouali delves into the utilization of kernel-based learning techniques alongside feature selection analysis within the realm of cancer diagnosis. The literature review underscores the importance of accurate and early cancer diagnosis, highlighting the limitations of traditional methods and the increasing interest in machine learning for diagnosis improvement. The study investigates how kernel-based learning effectively manages intricate medical data, examining diverse kernel functions and their relevance to cancer diagnosis. Additionally, it emphasizes the significance of feature selection in enhancing model performance and interpretability, addressing challenges like the curse of dimensionality. In summary, the review provides a comprehensive grasp of kernel-based learning and feature selection within cancer diagnosis, establishing the foundation for the paper's aim of enhancing diagnostic models.

• Chenyang Liu1, Shen-Chiang Hu1, Chunhao Wang2, Kyle Lafata2, Fang-Fang Yin1,2, "Automatic Detection of Pulmonary Nodules on CT Images with YOLOv3: The study "Development and Evaluation Using Simulated and Patient Data" scrutinizes existing literature concerning automatic pulmonary nodule detection on CT images, with a focus on implementing deep learning techniques such as YOLOv3.The literature review emphasises the value of early lung nodule detection as well as the difficulties in using manual detection techniques. The discussion revolves around advancements in deep learning for analysing medical images, particularly highlighting the effectiveness of CNNs in accurately detecting anomalies such as nodules. It also discusses the drawbacks of conventional detection methods as well as YOLOv3's advantages, like enhanced accuracy and real-time processing. It also examines earlier research that used deep learning—YOLOv3, among others—for nodule diagnosis on CT scans, going over methods and performance measures. All things considered, the review provides a thorough summary of automatic pulmonary nodule identification methods, setting the stage for the investigation in the study.

• Preeti Joon, Shalini Bhaskar Bajaj, and Aman "Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Techniques" provides an extensive overview of research findings regarding the segmentation and detection of lung cancer utilizing image processing and clustering methodologies. The literature review highlights the potential of image processing approaches to improve detection accuracy and highlights the limits of standard diagnostic methods, emphasising the significance of early and accurate lung cancer identification for better patient outcomes. It addresses issues such as noise and structural differences and covers a variety of image processing approaches, with a particular emphasis on segmentation strategies to extract lung areas in CT images. It also discusses the application of clustering algorithms, Highlighting advantages like categorization according to texture or intensity information, to pinpoint areas of concern within segmented images that could indicate the presence of lung cancer.. In addition, previous research integrating image processing and clustering for lung cancer diagnosis is reviewed, with methods, datasets, and performance measures described. In summary, it provides a thorough outline of studies on lung cancer segmentation and detection, which establishes the foundation for the paper's contributions.

• In-depth analysis of previous studies pertinent to the use of spectral reflectance and machine learning algorithms for the detection of pepper fusarium disease is provided in "Detection of Pepper Fusarium Disease Using Machine Learning Algorithms Based on Spectral Reflectance" by Kerim Karadağ a, Mehmet Emin Tenekeci b, Ramazan Taşaltın a, and Ayşin Bilgili c.

The literature review addresses issues with conventional detection techniques while highlighting the significance of early identification of plant diseases like Fusarium to maintain crop output and quality. It emphasises how spectral reflectance analysis can be used for early and non-destructive detection.

The writers go over earlier research that used spectrum reflectance methods to identify diseases in a variety of crops, including peppers. Additionally, the discussion delves into methods such as spectroscopy and hyperspectral imaging.

Moreover, the research investigates the application of machine learning algorithms in analysing spectral reflectance data to detect diseases, emphasizing the advantages of these methods and examining prior studies that have utilized algorithms such as SVM, random forests, and neural networks. In summary, the review provides a thorough overview of the literature on the use of spectral reflectance and machine learning methods in pepper Fusarium disease diagnosis, thereby laying the groundwork for the study presented in this paper.

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

As of my last update in January 2022, machine learning algorithms and data analysis techniques have become prevalent in lung cancer prediction systems. These systems leverage various patient data sources, including genetic profiles, imaging diagnostics, and medical histories, to identify patterns and risk indicators associated with lung cancer. Continuous refinement of algorithms through the analysis of historical data enhances prediction accuracy over time. Moreover, advancements in deep learning and image recognition hold promise for further enhancing the interpretation of medical images like CT scans within these systems. While these current methodologies represent significant progress in early detection, ongoing research and development are crucial to advancing their capabilities and incorporating the latest advancements in medical science.

3.2 PROPOSED SYSTEM

The proposed system for a mini-project on lung cancer prediction involves constructing a robust machine learning model that incorporates relevant factors such as patient demographics, medical history, genetic data, and potentially imaging data. By employing algorithms to analyse and detect patterns in historical records, the system aims to predict lung cancer risk accurately. Ensuring data quality and model accuracy will require integration with electronic health records and advanced data preprocessing techniques. The system's user-friendly interface facilitates early identification and intervention by enabling healthcare providers to input patient information and receive real-time forecasts. Continuous updates and refinements to the model are essential to enhance predictive accuracy and reliability over time. Ultimately, the system aims to equip medical professionals with an effective tool for proactive lung cancer risk assessment, fostering prompt preventive measures and personalized patient care.

IV. SYSTEM SPECIFICATION

4.1 HARDWARE REQUIREMENTS

• Processor	: Intel Core i3
• Ram	: 4GB
• Hard disk drive	: 1TB

4.2 SOFTWARE REQUIREMENTS

• Operating system	: Cross Platform
• IDE used	: Jupyter Notebook
• Documentation	: Microsoft Word

4.3 SOFTWARE DESCRIPTION

a. Front End: HTML

HTML, or Hypertext Markup Language, is the fundamental building block of web development, used to create the structure and layout of web pages. In any project involving HTML, whether it's a simple personal website, a complex web application, or anything in between, HTML serves as the backbone. HTML consists of a series of elements or tags that define the different parts of a webpage, such as headings, paragraphs, images, links, forms, and more. These elements are structured in a hierarchical manner, forming the layout and content of the webpage.

DOI: 10.17148/IARJSET.2024.115112

In a project, HTML is used to organize and structure the content, making it readable and accessible to users and search engines alike. HTML serves as the foundation for integrating other technologies such as CSS (Cascading Style Sheets) and JavaScript to improve a website's aesthetics and functionality. Its versatility enables developers to craft responsive designs, ensuring seamless user experiences across diverse screen sizes and devices. Moreover, HTML5, the latest iteration, introduces numerous features and APIs, enriching multimedia experiences, enhancing accessibility, and refining web content semantics. In essence, HTML forms the bedrock of every web Endeavor, offering the essential structure and coherence needed to develop compelling and operational web pages..

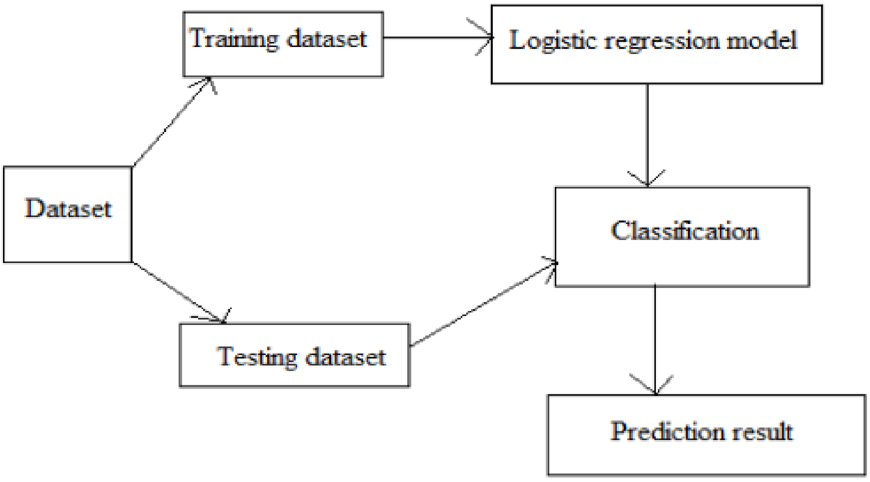
b. Back End: Machine Learning

Machine learning, a branch of artificial intelligence, enables computers to learn from data and make predictions or decisions without explicit programming. It's a transformative field with diverse applications spanning industries like healthcare, finance, marketing, and entertainment. Within project contexts, machine learning presents endless opportunities for tackling complex problems and extracting valuable insights from large datasets. At its core, a machine learning project comprises several pivotal stages: data collection, preprocessing, model selection and training, evaluation, and deployment. Each stage demands careful consideration and expertise to ensure project success. Data collection marks the initial phase, involving the gathering of pertinent datasets from varied sources like databases, APIs, or sensors. The quality and quantity of data significantly influence model performance, making preprocessing pivotal.

This phase entails data cleaning, handling missing values, outlier removal, and feature transformation to render them suitable for analysis. Following data preparation, the subsequent step is model selection and training. This necessitates choosing the appropriate machine learning algorithm based on the problem's nature and data characteristics. Common algorithms include linear regression for continuous outcome prediction, decision trees for classification, and neural networks for intricate patterns.

The chosen model is then trained on the data to grasp underlying patterns and relationships. Evaluation is crucial for assessing the trained model's performance and its generalization to unseen data. This involves dataset partitioning into training and testing sets or employing techniques like cross-validation. Metrics such as accuracy, precision, recall, and F1-score are typically used to gauge model performance and pinpoint areas for enhancement. Finally, deployment entails integrating the trained model into a production environment for real-time predictions or decisions. This often necessitates deploying the model as a web service or embedding it within existing systems. Continuous monitoring and updating are imperative to ensure the model remain accurate and pertinent over time.

4.4 METHODOLOGY



This figure represents the flow of the data in the classification process. Data are split for testing and training and then sent for model classification. Finally, the model is used for prediction.

DOI: 10.17148/IARJSET.2024.115112

a) COLLECTION OF DATA:
Most suitable input data has been taken; data can be found from any sources. In this research paper, the datasets are collected from online.

b) LUNG CANCER DATASET AS INPUT:
In this step, we give the dataset as an input to the proposed system and it gives the result. The dataset is split into two parts training and testing.

c) FEATURE SELECTION:
The aim of the feature selection is to identify those inputs, which are related with output values, and the values depend upon some input, which is chosen by using some test.

d) SPLITTING DATASET:
The dataset comprises 900 test results, which are divided into training and testing subsets. By utilizing these segments for training and evaluation, we can calculate the accuracy score of the model.

e) PERFORM LR TECHNIQUE ON TRAINING DATASETS:
Logistic regression is one of the popular models used for analysing many datasets in the machine learning. In this logistic regression is mainly used for classification purpose.

V. IMPLEMENTATION AND RESULT ANALYSIS

5.1 IMPLEMENTATION

To implement we need logistic regression equations. The lung cancer dataset included attributes hence the hypothesis for lung cancer detection is of the form.

Here x1, x2, x3, x15 are 15 attributes which facilitates lung cancer. Total number weights in this case are 16 including

The cost function for logistic regression is given by

where m is the total number of input instances.

The objective here is to minimize the cost function parameterized by 0, using gradient descent rule

Here partial derivates of cost function parameterized by using gradient descent rule ie...,

Pseudocode:

```
1: Initialize all weights θ 's to zero
2: Use equation (3) to estimate θTx
3: Obtain the hypothesis θ(x) from equation (2)
4: Compute the cost function θ (8) from equation (4)
5: Compute the gradients using equation (6) and equation (7). Update the weights θ 's using (5).
6: Go to step2, repeat the process until the weights do not change.
7: For predicting class of new data, optimal weights corresponding global minimum cost function is recorded and substituted in step3. If probability is above 0.5 lung cancer positive else lung cancer negative.
```

5.2 RESULT

In this study, we employ logistic regression to forecast the impact of lung cancer on patients. The dataset is partitioned into two segments, with 70% allocated for training and the remaining 30% for testing purposes. Through iterative training on various subsets of the training data, the system learns to predict lung cancer occurrences. Subsequently, it undergoes testing on the reserved dataset to ensure accuracy. Based on the obtained score, we ascertain whether an individual is likely to have lung cancer. The training score achieved by the model stands at an impressive 0.991736.

VI. CONCLUSION

Lung cancer stands as one of the leading causes of cancer-related deaths worldwide, both in terms of prevalence and fatality. Unfortunately, many individuals fail to seek timely treatment during the initial stages, leading to challenges in managing the disease in its advanced stages. Therefore, early precautionary measures significantly decrease the mortality rate associated with lung cancer. Machine learning techniques offer promising avenues for predicting lung cancer in its early phases. Implementing such systems can substantially reduce the mortality rates among patients.

REFERENCES

[1]. Guruprasad Bhat, Vidyadevi G Biradar, H Sarojadevi Nalini, (2012), “Artificial Neural Network based Cancer Cell Classification (ANN – C3)”, Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.

[2]. Mohamad Sayed, “Biometric Gait Recognition based on machine learning algorithms”. Journal of Computer Science, vol. 14(7), pp.1064 – 1073. DOI: 10.3844/jcssp.2018.1064.1073, 2018.

[3]. https://arxiv.org/pdf/1803.08375.pdf

[4]. Wu Liu and Cheng Zhang, 2018, ‘Learning Efficient spatial-temporal gait features with deep learning for human identification’, Springer Neuro Informatics, pp. 457–471, doi:10.1007/s12021-018-9362-4

[5]. Privietha P, Joseph Raj V (2022), "Hybrid Activation Function in Deep Learning for Gait Analysis," 2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON), Chennai, India, 2022, pp. 1-7, Doi: https://doi.org/10.1109/PECCON55017.2022.9851128.