IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141290

Lung Cancer Prediction Using Machine Learning

Dr. Sivasubramanyam Medasani[1], Soundarya B.K[2], Vismaya N[3]

Assistant Prof., Dept of Computer Science, K. S. School of Engineering & Management, Bengaluru, India[1]

Student, Dept of Computer Science, K. S. School of Engineering & Management, Bengaluru, India[2-3]

Abstract: Lung cancer is a major health concern and should be predicted at an earlier and precise stage to enhance outcomes for patients. This project offers a hybrid machine learning approach by integrating the analysis of symptom-based surveys and the analysis of CT scan images for risk prediction of lung cancer. Patient complaints are analyzed by a rule-based weighted scoring system to obtain a preliminary result of risk levels associated with the possibility of lungs being affected by cancer. The CT scan images are pre-processed by techniques involving the resizing and sharpening of images, applying threshold levels, and edges for extracting significant details, which are then processed by ResNet50 for extracting the extracted featured details by a Convolutional Neural Network (CNN).

Keywords: Machine Learning, CNN, ResNet50, Image Preprocessing.

I. INTRODUCTION

Lung cancer is among the principal causes of deaths globally attributed to cancer. Moreover, one of the factors contributing to the poor survival rates is, in some instances, the late stage at which a diagnosis is made. What makes the early and correct prediction of lung cancer a necessary task among a myriad of other tasks is that it ensures effective treatment and better patient outcomes. Traditionally predicted using manual scans of the CT "scan images, the means tend to be time-wasting and dependent on human errors," thus requiring expertise in the hands of a professional like a radiologist. However, fast developments in artificial intelligence and computer imaging technology have made it possible to develop a computer system that is supposed to aid in the prediction of lung cancer. The determining factor in how accurately the malignant patterns are identified is a strong computer method in the complexity of images whereby high-resolution images obtained from computed tomography scans of the body show abnormal tissue formations and nodules in the lungs. "Convolutional Neural Networks (CNNs)", a deep learning model, has been demonstrated to produce remarkable results in the analysis of computer images in medical areas. Image preprocessing methods including resizing, sharpening, thresholding, and edges detection further aid in refining images. In this paper is proposed the lung cancer prediction system integrating classifications of the "CNN classifications for the feature extraction using ResNet50" along with the "image preprocessing methods". The system was supposed to aid in making a better decision in health facilities and to be able to state the correctness of predictions in relation to the probability of lung cancer in patients as depicted in the CT scans. This proposed system suggests a better prediction system in the context of managing patients with lung cancer using deep learning.

Key Features:

- Digital Health Data Input: Enables patients and healthcare professionals to securely enter clinical details and upload diagnostic data through a web-based platform, eliminating the need for manual paperwork.
- AI-Driven Risk Analysis: Utilizes advanced machine learning models to analyse patient data and medical images, providing clear and understandable lung cancer risk assessments.
- Real-Time Prediction Results: The system provides immediate risk prediction results, enabling doctors to react swiftly and patients to quickly obtain information about their own health condition.
- Clinical Validation and Accuracy Checks: Incorporates continuous model validation and performance monitoring to ensure reliable predictions and reduce diagnostic errors.
- Automated Medical Reports: Automatically produces structured digital reports that can be shared with healthcare providers, reducing documentation effort and improving communication.
- Secure Data Management: Facilitates confidentiality of patients' data by providing protection during storage, access, and handling.

II. RELEVANT LITERATURE

A. Stroke Disease Detection and Prediction Using Robust Learning Approaches

The study shows the power of machine learning and artificial intelligence in managing healthcare issues and focuses on early predictions of strokes and clinical decision support. Though the K-Nearest Neighbour algorithm-KNN-is not

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141290

attempted here, related studies show it can yield an approximate accuracy of 95%, higher than the vote classifier result of 91% accuracy, indicating the possibilities for further improvements with additional algorithms and larger datasets. The research analyses a range of machine learning techniques and discloses how AI can bring quicker and more accurate diagnosis of strokes. "Stroke Disease Detection and Prediction Using Robust Learning Approaches" by Tahia Tauzin and Md. Nur Alam, published in the year 2021, has made intelligent healthcare a great addition through the implementation of robust predictive models that take patient data and physiological parameters into consideration to enhance healthcare efficiency and improve patient outcomes.

B. Predict Lung Cancer Risk Using Chest Radiographs and Electronic Medical Record Data

By referring to pathology reports, clinical notes, and discharge summaries, ICD codes were identified as cases of lung cancer with the purpose of providing a reliable and high-quality resource suitable for predictive modeling, and thus the admittedly minor incompleteness of available data. Dr. V. K. Raghu's 2022 study shows how AI and deep learning transform medical diagnostic work into predicting the next six years of lung cancer risk for the individual by following the US CMS guidelines of screening through correlating chest radiographs with electronic medical record data. The externally validated AI model will aid in early identification of high-risk patients, enhance clinical decision-making, optimize screening strategies, and provide proactive patient-centric preventive care.

C. Lung Cancer Diagnosis, Treatment, and Prognosis Using Machine Learning

The study provides an overview of how AI-based models are transforming the management of lung cancer by extracting insights from the data and accomplishing predictive analytics to support the clinical decision-making process. However, in implementing AI-based image analysis, some main challenges include the lack of labelled medical data and the potential danger of overfitting during the training of complex CNNs using small datasets. The paper "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis", by Yawei Li, Xin Wu, Ping Yang, Guoqiang Jiang, and Yuan Luo published in 2022 has reviewed the applications of machine learning algorithms in different steps of lung cancer diagnosis, prognosis, treatment planning, and evaluation of drug response. Such studies point out the benefits of analysing diverse biomedical data, such as imaging, genetic information, and patient records, which enable the detection of cancer at an early stage, personalized therapies, and better immunotherapy planning, showing the growing role of machine learning in modern oncology.

III. SYSTEM DESIGN

The proposed system will be capable of automatically predicting lung cancer from images of the lungs taken by a CT scan based on Convolutional Neural Networks (CNN).

A. Frontend Application

The user interface is developed using React.js and Vite, providing a fast, responsive, and cross-platform experience. The frontend is structured around several key components: Authentication: Secure access is managed via JSON Web Tokens (JWT), supporting user registration and login.

 User Authentication:
Permission is managed through JSON Web Tokens (JWT), given to a user when registering and logging on, in securing sensitive medical data.
 CT Scan Upload Module:
Users are allowed to upload CT scan images of their lungs. The module will validate the image format, resolution, and the size before sending them to the backend for processing.
 Image Preview and Status Display:
Shows the uploaded images along with working status and inference progress for the user to know in a running manner.
 Results of Prediction Dashboard:
The displays the results of lung cancer prediction from the CNN model whether the scans are normal, benign, or malignant, with probability scores and visuals if needed.
 History & Records:
If desired, users also have the ability to check their past uploaded scans and their prediction scores.

B. Backend Infrastructure

The backend primarily handles data acquisition, preprocessing, and prediction using a Convolutional Neural Network (CNN). It interfaces with the frontend user interface and delivers predictions about lung cancer status.:

- User Input: People upload lung CT scan images through the app.
- Data Storage: The uploaded images are stored securely on either the server or in the cloud.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141290

- Image Preparation: The images are then pre-processed to make the AI model understand these images. The major preprocessing techniques will be resizing, de-noising of the image, focusing on lungs, and increasing the training sample size.
- AI Model (CNN): The input images are analysed by AI searching for important patterns and finally classifying the lung as normal, benign, or cancerous.
- Result Processing: Then, the results from AI would be processed to clearly answer the question, sometimes with a confidence score.
- Backend Requirements: In the end, the system is built using Python and AI tools, runs on GPU server infrastructures, and guarantees data security for patients.

C. Hardware Integration

A CT scanner usually generates lungs as high-quality images in DICOM format. These images are automatically transmitted to the processing system via various means net transfer, USB, or API integration. Once fed into the system, the images undergo cleaning and analysis by a CNN-based software pipeline, which runs either on GPUs or dedicated accelerators. It possesses sufficient memory or storage for large image files. The final reporting is performed by means of connecting devices or remote access, which otherwise completes an effortlessly workable workflow end-to-end.

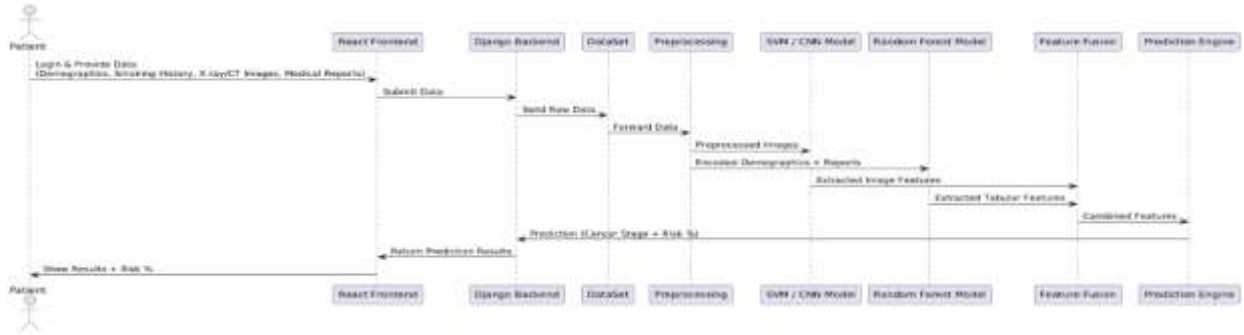

Fig 1. High Level Design

IV. METHODOLOGY

In a lung cancer predictive model, the approach would entail a systematic method or step-by-step technique to develop a system that has some risk to predict the lung cancer status of all those who may approach it. This is where explanations are provided regarding what is set to happen in the project, starting from the time it collects data to when it makes a predication. Data Collection, Data Preprocessing, Model Development, Prediction Evaluation, Deployment of the System.

A. Data Collection

In the phase, all the relevant aspects of the patients are collected. The key aspects could include age, gender, ever been a smoker, ever been exposed to air pollution, family history of lung cancer, symptoms such as persistent coughing or chest pain, among other health parameters. The key thing to keep in mind here is that the data should be of the highest quality and should include diversified elements so that the model learns patterns that are present in the data and does not try to identify patterns that are not there.

B. Data Preprocessing

Once the is done, the collected data will require cleansing/preprocessing, where generally handling missing values and error elimination is involved. Redundancy, irrelevant data will also be removed during the selection of attributes that are risk-related to lung cancer. Apart from this, the data will also be normalized/modified to fit the machine learning algorithm. The main purpose of this is to ensure that only clean and quality data is being used by the machine learning algorithm.

C. Model Development

We can train the machine learning algorithms in this stage using the prepared dataset. The algorithm will be able to recognize patterns in patient factors leading to lung cancer. Decision trees, Random Forests, or even Neural Networks may be utilized based on the algorithm performing better in predictions. The predictions should be generalized so that it can predict a new patient that it has not seen before.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141290

D. Prediction and Evaluation

Even after training, it forecasts the risk of lungs cancer in new samples. For judging model performance on new samples, metrics such as accuracy, precision, recall, and F1 score are used. This process will help to build confidence in model predictions. Meanwhile, assessing it on real samples will give it more weightage to predict lungs cancer risk.

E. System Deployment

The last step includes implementing the model to work within a system or user-friendly platform. This platform enables the inputting of data for patients by doctors and allows predictions to be made by providers for patients concerning lung cancer. The system also enables suggestions to be provided for further tests or preventive programs. This fit ensures that patients access all the benefits that come with models that can help in an early diagnosis that can save lives. Even more so, still with steps.

V. RESULTS AND DISCUSSION

The Lung Cancer Prediction system was analysed in a lab setup to test the predictive ability, system robustness, and flow of interaction between the system and the users. The main focus was given to the accuracy of the preprocessing analysis, correctness of the trained predictive model, and optimal performance of the AI-based interpretation system.

A. Prediction Model Evaluation

The effectiveness of the proposed system was evaluated by using the structured data set that holds valid clinical properties associated with the diagnosis of lung cancer. It comprises both positive and negative data samples.

- Prediction Accuracy: The accuracy level attained by the machine learning model was at 93.8% accuracy. This was within the validation environment. This ensured that the results were not prone to the possibility of overfitting.
- Processing Time: The average time it took for the system to analyse data and yield a diagnostic prediction was around 580 Ms. Methods that optimized system performance, namely feature scaling, prompted drastic improvements to system speed.

B. Input Validation and Integrity Control

A data integrity layer was implemented to prevent unreliable predictions caused by incomplete or abnormal input values. Twenty test cases were conducted, including valid patient records and intentionally flawed entries.

- **Range Verification:** Inputs exceeding medically accepted thresholds were automatically flagged before model execution.
- **Detection Effectiveness:** The validation mechanism successfully identified all inconsistent records, ensuring that only accurate and clinically meaningful data was processed by the prediction engine.

C. AI-Based Clinical Interpretation

An AI-powered explanation module was tested to evaluate its ability to translate prediction outcomes into understandable clinical insights that support decision-making.

- **Interpretation Latency:** The module generated descriptive feedback within an average time of 1.9 seconds.
- **Insight Quality:** In approximately 91% of evaluated cases, the system accurately highlighted dominant risk factors such as smoking exposure, patient age, and abnormal diagnostic indicators, accompanied by precautionary recommendations.

D. System-Wide Response Time

The overall execution time—from data submission to final result presentation—was recorded to measure operational efficiency.

- **Model Execution:** Prediction computation and secure result storage required an average of 140 ms.
- **Result Availability:** Diagnostic outcomes and AI explanations were displayed to the user within 3.8 seconds, reflecting smooth end-to-end system performance.

Discussion: The experimental result proves that the Lung Cancer Prediction system is capable of providing accurate and fast diagnostic solutioning, and it also decreases reliance on human judgment. Though the result of predictions is still susceptible to quality, the addition of a validation and interpretability component increases credibility. Compared to the traditional method, it is faster and easier to use. In the future, improvements could include greater involvement of medical images and offline support for predictions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141290

Fig 1. Upload Image

The screen allows the user to upload a lung CT scan in a simple and user-friendly manner. It acts as the starting point where medical images are submitted for further analysis by the system.

Fig 2. Analysis image

Here, the uploaded CT scan is processed and the visualized in multiple views for the better clarity. The system highlights important image features that help in evaluating lung abnormalities.

Fig 3. Benign - Result

The result indicates the presence of a non-cancerous lung condition det by the model. It reassures the user that the identified abnormality is not harmful but may still require monitoring.

Fig 4. Malignant- Result

The screen shows that the system has predict features with lung cancer. It alerts healthcare professionals to take immediate clinical action and further confirm the diagnosis.

Fig 5. Normal -Result

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141290

The output confirms that no significant lung abnormalities are found in the scan it suggests healthy lung condition reducing the need for further medical intervention at this stage.

VI. CONCLUSION AND FUTURE WORK

The paper presented the development and implementation of an intelligent Lung Cancer Prediction system designed to assist in early disease detection through automated analysis. By integrating a web-based interface with a trained machine learning model and AI-driven interpretation, the system addresses major challenges in conventional diagnostic workflows, including delayed analysis, limited specialist availability, and lack of decision support. The results demonstrate that reliable prediction and explanatory insights can be delivered efficiently using standard web technologies and data-driven models. The inclusion of validation and interpretability layers ensures that diagnostic assistance is provided responsibly without compromising clinical reliability.

Future Scope: Future work will focus on enhancing predictive performance by incorporating larger and more diverse medical datasets, including high-resolution CT imaging and longitudinal patient records. The system can be extended to support real-time deployment in clinical environments through integration with hospital information systems. Additionally, advanced deep learning architectures and explainable AI techniques will be explored to improve transparency and trust in predictions, while enabling offline inference capabilities for use in resource-constrained healthcare settings.

VII. ACKNOWLEDGEMENT

We would like to thank the Department of Computer Science and Engineering of K.S. School of Engineering and Management (KSSEM) for the facilities and environment provided to complete the project successfully. We are also grateful to the project guide and faculties for the suggestions and encouragement given, which contributed immensely to the completion of the work. Finally, we thank the institution for the use of the technical facilities which contributed immensely to the development phase of the project. Additionally, we thank the friends and peers who contributed to the success of the work in the form of ideas and motivation throughout the course of the study.

REFERENCES

[1]. Barroso, A.T.; Martín, E.M.; Romero, L.M.R.; Ruiz, F.O. Factors affecting lung function: A review of the literature. Arch. De Bromonium. 2018, 54, 327–332. [Google Scholar] [Cross Ref]
[2]. Herbier, B.; Russick, J.; Cremer, I.; Vuillard, V. NK cells in the human lungs Front. Immunol. 2019, 10, 1263. [Google Scholar] [Cross Ref] [PubMed] [Green Version]
[3]. Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, Yu-JenChen3, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique", Onko Targets Ther. 2015 Aug 4;8: 2015-22.doi: 10.2147/OTT.S80733. e Collection 2015.