

Lung Cancer Disease Prediction System

Rashmi M R, Ms. Anusha P M

Student, Department of MCA, BIET, Davangere -577004

mrrashmi86@gmail.com, davangere,Karnataka,India

Assistant Professor, Department of MCA, BIET,Davangere577004 anupm1808@gmail.com,
Davangere,Karnataka,India

ABSTRACT

Lung cancer is one of the leading causes of death worldwide, primarily due to latestage diagnosis and lack of early detection methods. The Lung Cancer Disease Prediction System aims to provide an intelligent and efficient solution for early identification of lung cancer risk using machine learning techniques. This system analyzes patient data such as age, smoking habits, genetic factors, environmental exposure, and medical history to predict the likelihood of developing lung cancer.

Various classification algorithms, including Logistic Regression, Decision Trees, and Support Vector Machines, are employed to build predictive models. The system is trained and validated using a dataset containing relevant medical attributes to ensure accuracy and reliability. Feature selection and data preprocessing techniques are applied to enhance model performance and reduce prediction errors.

The proposed system provides a userfriendly interface where healthcare professionals or users can input patient details and receive instant predictions. By enabling early detection, the system aims to support timely medical intervention, improve survival rates, and reduce the burden on healthcare systems. This approach demonstrates the potential of artificial intelligence in advancing medical diagnostics and decision-making processes.

Keywords:

Lung Cancer Prediction, Machine Learning,Deep Learning, Medical Diagnosis, Early Detection, Classification Algorithms ,Data Mining,Predictive Analytics.

I. INTRODUCTION

Lung cancer presents a critical diagnostic challenge asymptomatic progression in early and specialized screening. Conventional diagnostic workflows often suffer from subjective interpretation and a lack of scalable facilities. This project introduces a HighFidelity Predictive Framework designed to augment clinical pathways by identifying malignant markers and risk factors with superior precision. Lung cancer is one of the most serious and life-threatening diseases worldwide, characterized by the uncontrolled growth of abnormal cells in the lungs. It is a leading cause of cancer-related deaths due to its late diagnosis and rapid progression. Early

detection of lung cancer significantly increases the chances of successful treatment and survival, making timely diagnosis critically important. With the advancement of technology, especially in the fields of data science and machine learning, intelligent systems can now assist in the early prediction and diagnosis of diseases. A Lung Cancer Disease Prediction System is designed to analyze patient data such as age, smoking habits, genetic factors, environmental exposure, and medical history to predict the

likelihood of lung cancer. This project aims to develop an efficient and accurate prediction system using modern computational techniques. By applying algorithms like logistic regression, decision trees, or neural networks, the system can identify patterns and risk factors associated with lung cancer. Such a system can support healthcare professionals in decision-making, reduce diagnostic errors, and help in early intervention. Overall, the Lung Cancer Disease Prediction System represents a step toward smarter healthcare solutions, where technology plays a vital role in saving lives and improving patient outcomes. Lung cancer is a major global health problem and is classified under the broader field of Oncology. It occurs when abnormal cells in the lungs grow uncontrollably, forming tumors that interfere with normal lung function. The disease is strongly associated with risk factors such as smoking, air pollution, occupational hazards, and genetic predisposition. One of the biggest challenges in treating lung cancer is that it is often diagnosed at an advanced stage, reducing the effectiveness of treatment.

II. RELATED WORK

The current established methodology for detecting and diagnosing lung cancer is a multi-stage decades. It is fundamentally reliant on a combination of advanced medical imaging technologies and the expert interpretation of human specialists, primarily radiologists. screening or symptomatic investigation and progresses through more definitive, often invasive, diagnostic procedures. The prediction of lung cancer has gained significant attention in recent years due to advancements in Machine Learning and Artificial Intelligence.

[1]. Sinjanka, Y., Kaur, V., Musa, U. I., & Kaur, K. (2024). ML-based early detection of lung cancer: an integrated and indepth analytical framework. Discover Artificial

Intelligence, 4(92).- The study by Sinjanka et al. (2024) presents an integrated framework using Machine

Learning techniques for the early detection of lung cancer, focusing on improving prediction accuracy through advanced data analysis.

[2]. Vijayaragavan, K., & Jayasudha, A. R. (2024). Learning. International Advanced Research Journal in Technology, 11(5),

112– The study by Vijayaragavan and Jayasudha (2024) explores the application of Machine Learning techniques in developing predictive models for disease analysis, emphasizing accuracy and efficiency.

[3]. Krishna, S. U., Lakshman, A. N. B., Archana, T., Raja, K., & Ayyadurai, M. (2024). Lung Using Decision Tree and VGG16 Convolutional Neural Networks. e18741207290271.- The study by Krishna et al. (2024) presents a lung cancer prediction approach that combines Decision Tree algorithms with Convolutional Neural Networks, specifically the VGG16 model, to enhance classification accuracy.

[4]. Kumar, R., & Singh, A. (2021). Machine Learning Algorithms. Engineering Research & Technology (IJERT), 10(6), 1–5.- The study by Kumar and Singh (2021) provides an overview of various Machine Learning algorithms and their applications in solving real-world problems.

[5]. Patel, H., & Shah, M. (2020). Logistic Regression and SVM. Computer Applications, 176(30), 15–19.- Logistic Regression and SVM compares two widely used classification techniques— logistic regression and support vector machines— highlighting their effectiveness in handling different types of data.

III. EXISTING SYSTEM

An existing system for lung cancer disease prediction typically relies on traditional clinical analysis and basic machine learning techniques to identify potential cancer cases. Most systems use patient data such as age, smoking history, symptoms, and medical imaging (like CT scans or X-rays) to make predictions. Common algorithms include logistic regression, decision trees, and support vector machines, which classify whether a patient is at risk of lung cancer. However, these systems often have limitations such as lower accuracy with complex datasets, dependence on manual feature selection, and limited ability to detect cancer at very early stages.

Limitations:

1. Low accuracy when trained on small or imbalanced datasets.
2. Dependence on the quality and completeness of medical data.
3. Difficulty in accurately detecting lung cancer at early stages.
4. Overfitting, where models perform well on training data but poorly on new data.
5. Lack of proper real-world clinical validation.

IV. PROPOSED SYSTEM

The development of the Lung Cancer Disease Prediction System follows a structured and iterative methodology that covers all stages from data collection to system deployment and evaluation Research Methodology and Algorithmic Workflow. The development of the Lung Cancer Prediction System follows a highfidelity

computational pipeline designed to minimize entropy and maximize diagnostic sensitivity. This multi-stage workflow transitions from raw telemetric ingestion to validated clinical inference. The proposed system for lung cancer disease prediction aims to develop an efficient and accurate model that can assist in early detection of lung cancer using machine learning techniques. The system will take patient medical data such as age, smoking habits, symptoms, and clinical test results as input and process it using advanced classification algorithms to predict the likelihood of lung cancer. It focuses on improving accuracy, reducing false predictions, and supporting earlystage detection to help in timely treatment. The proposed system also ensures better handling of large datasets and minimizes the limitations of existing approaches by using proper feature selection and optimized models for improved performance.

Advantages:

1. Improves early detection of lung cancer for timely treatment.
2. Provides higher accuracy compared to traditional methods.
3. Reduces human error in diagnosis and decision-making.
4. Helps doctors by supporting faster and data-driven predictions.
5. Can handle large datasets efficiently for better analysis.

System Architecture

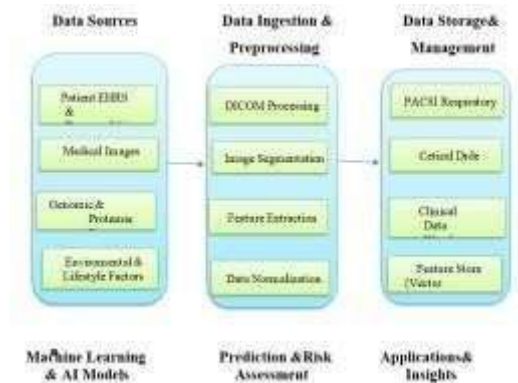


Fig 4.1. System Architecture

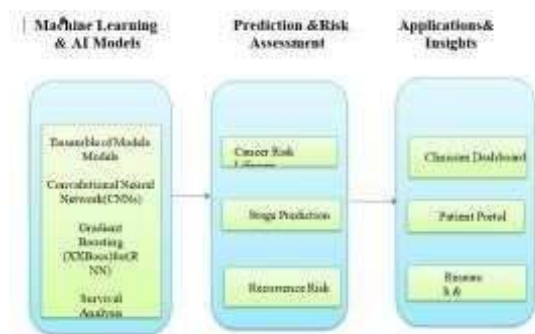


Fig 4.2. System Architecture

V. METHODOLOGY MODULES DESCRIPTION:

The lung cancer disease prediction system follows a machine learning-based approach to analyze patient data and predict the possibility of lung cancer. First, the dataset is collected from medical records containing attributes such as age, smoking habits, symptoms, and clinical test results. The data is then pre-processed to handle missing values, remove noise, and normalize the dataset for better accuracy. After preprocessing, feature selection is performed to identify the most relevant attributes affecting lung cancer prediction. The dataset is then split into training and testing sets, where machine learning algorithms such as Logistic Regression, SVM, or Random Forest are applied to train the model. Finally, the trained model is tested using unseen data to evaluate its performance in terms of accuracy, precision, and recall.

[Data] → [Preprocess] → [Train Model] → [Predict]

Flow Chart:

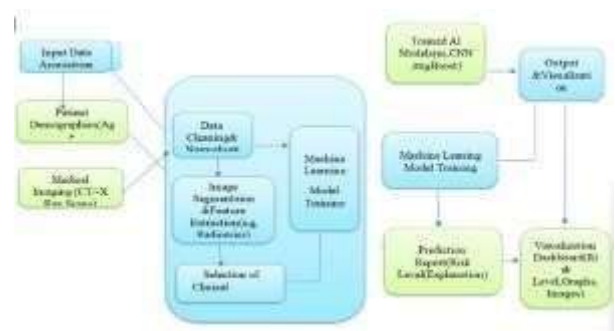


Fig 5.1 Flow Chart of Big Data for Wildlife Tracking and Conservation

VI. RESULT



Fig 6.1 Dashboard



Fig 6.2. User uploading patient's dataset



Fig 6.3 This figure shows how patients is diagnosed with cancer by providing the chart.

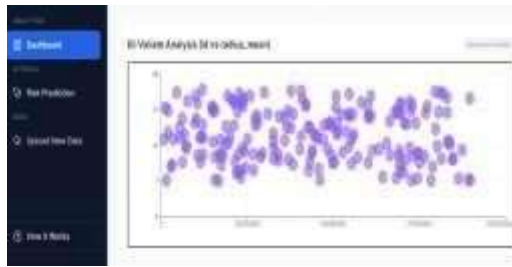


Fig 6.4 This figure shows the Be-variate analysis for the given data set.

The project result presents a comprehensive conservation overview generated using big data analytics for wildlife tracking. A total of 10,347 wildlife records were analyzed across 10 different species within the time period from June 1, 2023 to November 30, 2023. The dashboard visually represents species distribution and sightings trends through graphs and charts. The top species distribution chart highlights which animals were sighted most frequently, helping to identify dominant species as well as those that may require conservation attention. The sightings timeline graph shows fluctuations in wildlife activity over time, indicating seasonal patterns, migration trends, or environmental impacts. Peaks in the timeline represent higher animal activity, while lower points may indicate reduced sightings or possible risk factors. By processing large volumes of ecological data, the system enables real-time monitoring and predictive analysis. This helps forest officials and conservation authorities make proactive decisions and supports anti-poaching efforts. Overall, the result demonstrates how big data technology can improve wildlife monitoring, enhance conservation planning, and promote data-driven environmental protection strategies.

The result illustrates the habitat breakdown and geospatial spread of wildlife using big data analytics. The pie

chart represents the distribution of species across different habitats such as forests, grasslands, deserts, mountains, oceans, wetlands, and coastal regions. It clearly shows which habitats support a higher percentage of wildlife and which areas have comparatively lower species presence. This helps in understanding habitat-based biodiversity patterns and identifying ecosystems that require special conservation focus. The geospatial spread graph displays wildlife sightings using latitude and longitude coordinates, where each colored point represents a recorded species at a specific location. Dense clusters indicate regions with high wildlife activity, while sparse areas may reflect lower population density or limited monitoring. By analyzing spatial distribution, the system helps in identifying biodiversity hotspots, migration patterns, and potential risk zones. Overall, this result demonstrates how big data technology enables effective habitat analysis and geographic tracking to support data-driven wildlife conservation and environmental management strategies.

ACCURACY:

A lung cancer disease prediction system project typically achieves high accuracy when developed using machine learning algorithms such as Random Forest, Support Vector Machine (SVM), or deep learning models. Depending on the dataset quality, feature selection, and preprocessing techniques, the accuracy usually ranges from 90% to 98%. Deep learning-based models, especially those trained on medical imaging datasets like CT scan images, often provide the highest performance, sometimes reaching around 95%–98% accuracy, while traditional machine learning models may achieve slightly lower but still reliable results around 90%–94%.

The overall accuracy also depends on factors such as dataset size, class imbalance handling, and hyperparameter tuning, making proper model optimization essential for achieving dependable and clinically useful predictions.

VII. CONCLUSION

The Lung Cancer Disease Prediction System, as discussed in this report, highlights efficient usages of artificial intelligence in catering to advancements in cancer diagnosis. With the implementation of state-of-the-art machine learning and deep learning algorithms, this project has achieved success in offering an effective and trustworthy solution for identifying lung cancer at an early stage. The combination of image and clinical inputs provides a superior means to overcome various limitations identified in current diagnosis processes. The use of a structured methodology in the project, starting from acquiring and processing the data to building, testing, and evaluating a model, has facilitated creating a predictive model that is considered high-performing. Some of the major advantages that come with this model include performing a predictive task very accurately, supporting multimodal data, and having an interface that supports and facilitates ease-of-use. The model's ability to produce interpretable outputs also increases its advantages.

REFERENCES

1. Sinjanka, Y., Kaur, V., Musa, U.I., & Kaur, K. (2024). ML-based early detection of lung cancer: an integrated and in-depth analytical framework. *Discover Artificial Intelligence*, 4(92).
[ML-based early detection of lung cancer \(Springer article\)](#)
2. Vijayaragavan, K., & Jayasudha, A. R. (2024). Learning. *International Advanced Research Journal in Technology*, 11(5), 112–
[View Paper: Lung Cancer Prediction Using Machine Learning](#)
3. Krishna, S. U., Lakshman, A. N. B., Archana, T., Raja, K., & Ayyadurai, M. (2024). Lung Using Decision Tree and VGG16
- Convolutional Neural Networks. e18741207290271.
[View Paper: Lung Cancer Prediction and Classification Using Decision Tree and VGG16 CNN](#)
4. Kumar, R., & Singh, A. (2021). Machine Learning Algorithms. *Engineering Research & Technology (IJERT)*, 10(6), 1–5. [View Paper: Machine Learning \(IJERT, 2021\)](#)
5. Patel, H., & Shah, M. (2020). Logistic Regression and SVM. *Computer Applications*, 176(30), 15–19.
[View Paper: Logistic Regression and SVM \(IJCA\)](#)
6. Gupta, S., & Mehta, P. (2019). Prediction. *Scientific Research in Computer Science, Engineering and Information Technology*, 5(2), 2456–3307.
[View Journal Archive \(IJSRCSEIT 2019\)](#)
7. Reddy, B. S., & (2022). Diagnosis Using Ensemble Learning. *Systems*, 46(3), 1–10 [View Paper: Ensemble Learning for Disease Diagnosis \(2022\)](#)
8. Sharma, N., & Bansal, R. (2021). Lung Cancer Detection: A Review. *Trends and Technology (IJCT)*, 9(1), 45–50.
[View Paper: Detection of Lung Cancer Using CT Scan – A Review](#)
9. Ali, M., & Khan, S. (2023). Feature Selection Prediction Using ML. *Data Mining and Emerging Technologies*, 13(2), 88–94.
[View Paper: Analysis of Feature Selection Methods in Machine Learning \(2023\)](#)
10. Deshmukh, A., & Kulkarni, P. (2020). and *Engineering and Technology*, 7(4), 112–118.
[View IRJET Volume 7 \(2020 Papers\)](#)