ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

Predicting Lung Cancer Using Machine Learning

Dr. A. Srinivasa Rao[1], (M.Tech, Ph.D.); Challa Manikanta[2]; Akuri Azad[3]; Kandagatla Rushikesh[4]; Polavarapu Rohitha[5]

[1]Computer Science and Engineering Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology Vijayawada, India
[2]Computer Science and Engineering Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology Vijayawada, India
[3]Computer Science and Engineering Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology Vijayawada, India
[4]Computer Science and Engineering Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology Vijayawada, India
[5]Computer Science and Engineering Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology Vijayawada, India

Publication Date: 2025/07/14

Abstract: Accurate and timely forecasting is essential for improving outcomes for patients because lung cancer is still one of the deadly illnesses in the world because of its late-stage detection and lack of effective early detection techniques. Conventional diagnostic methods frequently involve invasive procedures, and although medical imaging methods like CT scans and X-rays offer useful information they need to be interpreted by professionals and can occasionally result in incorrect diagnoses or postponed treatment. Furthermore, cancer of the lung risk is influenced by various variables including demographic variables (such as gender and age), daily behaviors (such as smoking) and signs of disease (such as persistent cough and other lung symptoms). This study uses machine learning approaches to create a strong lung cancer prediction model based on a large dataset that includes clinical, lifestyle and demographic characteristics in order to overcome these obstacles and improve predicted accuracy. To guarantee data quality and dependability prior to model training the dataset is subjected to comprehensive exploratory data analysis (EDA), preprocessing and feature scaling. Key performance metrics like accuracy, mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are used to implement and assess a variety of machine learning models, such as Deep Neural Networks (DNN), Decision Trees and Random Forests The Decision Tree and Random Forest models perform noticeably better with accuracies of 90.32% and 93.54%, respectively whereas the DNN model performs sub optimally, according to preliminary results, with an accuracy of 12.62%. The Bagging Classifier is used to further optimize the Random Forest model which has the best accuracy in order to improve performance and stability.

Keywords: *CNN, Data Augmentation, Disease Detection, Plant Health, Real-Time Detection and Streamlit.*

How to Cite: Dr. A. Srinivasa Rao; Challa Manikanta; Akuri Azad; Kandagatla Rushikesh; Polavarapu Rohitha (2025). Predicting Lung Cancer Using Machine Learning. *International Journal of Innovative Science and Research Technology*, (RISEM–2025), 87-96. https://doi.org/10.38124/ijisrt/25jun173

I. INTRODUCTION

Millions of people die from lung cancer every year making it one of the main causes of death from cancer globally. Due to the absence of obvious symptoms in the early stages the disease is frequently detected at an advanced stage which reduces the effectiveness of treatment and drastically lowers survival rates [1]. Traditionally, imaging technologies like computed tomography (CT) scans and chest X-rays as well as invasive diagnostic procedures like biopsies have been used to diagnose lung cancer. Despite advancements over time these methods' efficacy is frequently constrained by issues including misunderstandings, exorbitant prices and accessibility issues in many areas. Researchers are looking on more effective and non-invasive ways to detect lung cancer because of its rising incidence especially among smokers and people exposed to dangerous environmental factors [2]. Data-driven approaches to disease identification and diagnosis have been made possible by recent developments in artificial intelligence and machine

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

learning which have created new opportunities for predictive analysis in the healthcare industry.

Medical imaging, biomarker analysis and conventional clinical evaluations based on symptoms and medical history are the main techniques currently used to diagnose lung cancer. High-resolution pictures of lung tissues can be obtained with CT and PET scans but they need to be interpreted by qualified radiologists which leaves room for human error and inconsistency [3]. These imaging methods are also costly and not always available especially in underdeveloped areas. Another option is to use blood samples for biomarker-based examinations but they are not always accurate for early-stage detection. Additionally, risk assessment has been done using statistical models and traditional machine learning techniques although their effectiveness is frequently hampered by their small datasets and poor generalization across a variety of populations. Owing to these drawbacks an automated data-driven approach that can incorporate a variety of elements such as patient demographics, lifestyle choices, symptoms and clinical biomarkers is desperately needed in order to increase the accuracy of lung cancer predictions and aid in early detection initiatives [4].

This study suggests a machine learning-based lung tumor prediction model to overcome these obstacles. It uses clinical parameters (like coughing symptoms and blood test results), lifestyle factors (like smoking habits) and demographic data (like age and gender) to accurately classify lung cancer. To improve model performance the dataset is subjected to a thorough preprocessing procedure that includes feature engineering, scaling and exploratory data analysis (EDA). Performance metrics including accuracy, mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are used to implement and assess a variety of machine learning methods such as Deep Neural Networks (DNN), Decision Trees and Random Forest. The findings show that the Random Forest model has the best accuracy, which encourages additional Bagging Classifier modification to increase dependability [5]. By providing a non-invasive and easily accessible method for early lung cancer detection this discovery makes a substantial contribution to society. This technology decreases reliance on costly imaging methods and improves diagnostic accuracy by combining statistical models with biomarkers of blood and symptom analysis. In healthcare settings the real-time predicting tool can be used to help physicians make well-informed judgments, especially in isolated locations with little access to specialized medical services. Furthermore, this methodology enables people to proactively evaluate their risk of developing lung cancer which promotes prompt medical consultation and may increase survival rates [6]. By bridging the gap between early detection and potent treatment, the application of such artificial intelligence (AI)-driven approaches in the healthcare sector will ultimately lessen the burden of lung cancer-related fatalities worldwide.

II. LITERATURE SURVEY

Numerous previous studies have investigated the prediction of lung cancer using a variety of conventional and computational methods. Imaging-based techniques like computed tomography (CT) scans, positron emission tomography (PET) scans and chest X-rays have historically been crucial in the diagnosis of lung cancer. Although these methods have been frequently employed to identify anomalies in lung tissues not all medical facilities may have access to sophisticated imaging technology and skilled radiologists which are necessary for their success. Furthermore, the gold standard for establishing lung cancer is still biopsy-based diagnosis, which entails taking tissue samples for microscopic analysis [7]. However, this approach is time-consuming, intrusive, and unsuitable for widespread early screening. Researchers have looked into using statistical models and rule-based systems to evaluate patient symptoms and demographic factors in order to estimate the risk of lung cancer in order to overcome the shortcomings of these conventional methods [8]. However, these models frequently have poor accuracy because of their limited feature representation.

Using datasets that include patient history, pulmonary symptoms, and test findings, machine learning and the use of artificial intelligence have recently been applied to the prediction of lung cancer. In order to categorize lung cancer risk based on structured data, some research have used algorithms including logistic regression, support vector machines (SVM) and decision trees [9]. Additionally, to improve the accuracy of CT scan processing, deep learning models in particular, convolutional neural networks or CNNs have been used for image-based diagnosis. But for training, these models frequently need substantial computer resources and big annotated datasets. Additionally, several studies have looked into the use of blood biomarkers for lung cancer screening, combining machine learning techniques with laboratory test findings to enhance predictive performance [10]. Notwithstanding these developments, problems with feature selection, data imbalance and poor generalization across a range of patient groups persist in current systems, underscoring the need for a more thorough and reliable predictive model.

III. DATA COLLECTION & PREPROCESSING

Numerous characteristics linked to lung cancer risk are included in the dataset used in this investigation including clinical symptoms like coughing and chest pain lifestyle choices like smoking frequency and duration, and demographic characteristics like age and gender. Along with these characteristics the dataset also includes blood report indicators including white blood cell count, hemoglobin levels and other biochemical markers that offer important information about the risk of developing lung cancer. To ensure a diverse representation of people with and without lung cancer the data was gathered from publicly accessible medical datasets, records from hospitals and patient health reports.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

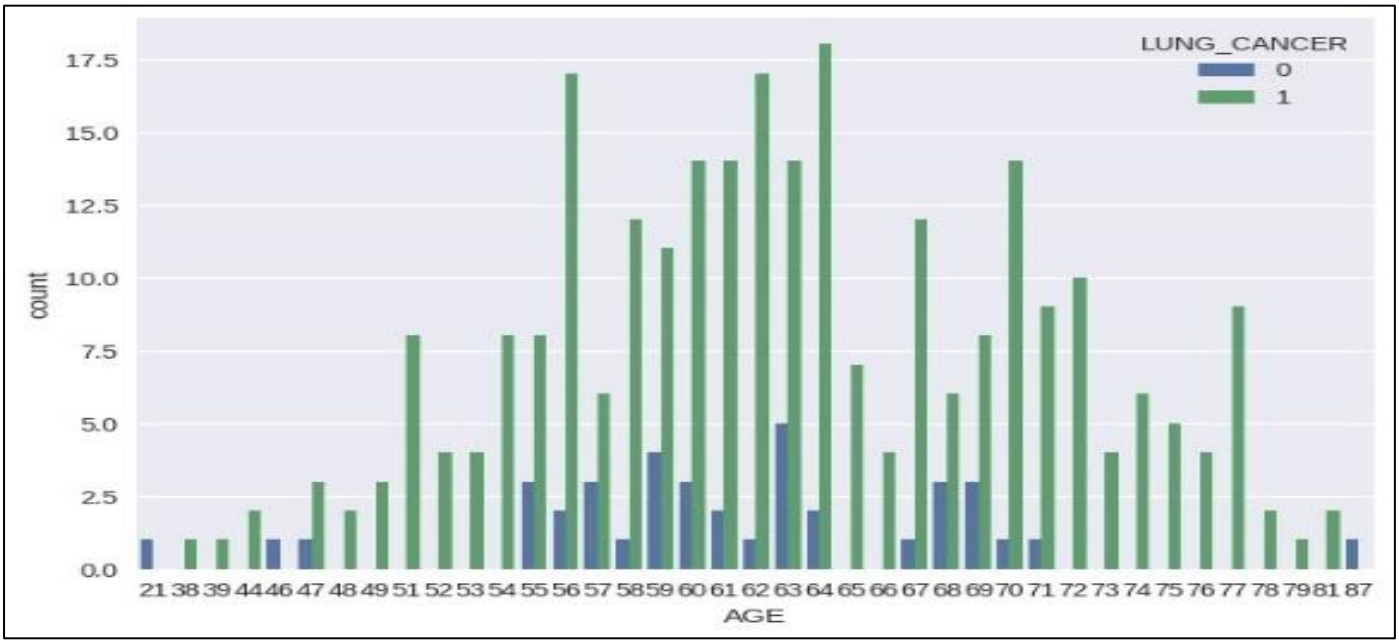

Fig 1 Analysis of Age vs Lung Cancer

Exploratory data analysis (EDA) was used to find and manage missing values, find anomalies and comprehend feature distribution in order to guarantee the dataset's dependability. Incomplete patient records or unreported symptoms frequently result in missing data in medical datasets [11]. Depending on the kind of missing values several imputation methods were utilized mode imputation was used for categorical features while mean and median imputation were used for numerical variables. Statistical techniques like z-score analysis and the interquartile range (IQR) approach were used to identify outliers in the dataset and the appropriate adjustments were made to lessen their influence on model performance.

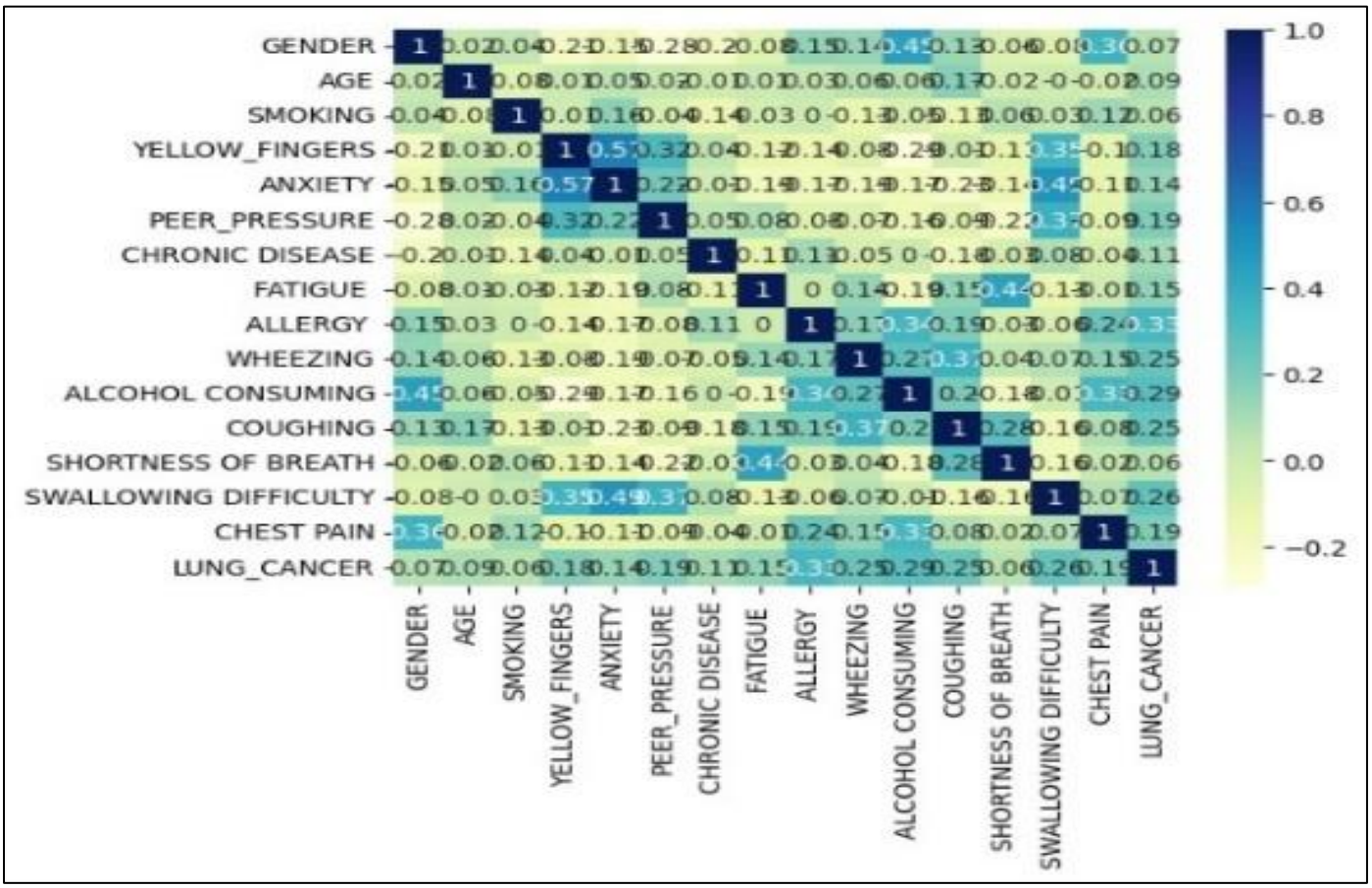

Fig 2 Correlation of Features in Dataset

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

In order to provide more useful qualities that could enhance the model's predictive power, feature engineering was utilized. To capture intricate interactions within the dataset, derived features were added including age-adjusted symptom severity scores a ratio of white blood cell count to hemoglobin levels and smoking duration. In order to transform categorical variables like smoking status and gender into numerical representations appropriate for machine learning algorithms, one-hot encoding and coding of labels approaches were used. In order to ensure that numerical characteristics remained within a consistent range and avoid dominance by qualities with bigger magnitudes the dataset was also standardized using the Min-Max Scaling and Standard Scaling techniques [12]. In order to decrease the quantity of input features while keeping the most pertinent data dimensionality reduction strategies like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were investigated. While RFE iteratively removed less significant features to increase model performance PCA assisted in capturing the variation within the dataset by converting associated characteristics into a set of orthogonal components [13]. By minimizing noise and redundancy these strategies were essential in maximizing the model's performance and making sure it learned from the most pertinent features. In order to balance the dataset in the event of a class imbalance and avoid biased predictions toward the dominant class the Synthetic Minority Over-sampling Technique (SMOTE) was also used.

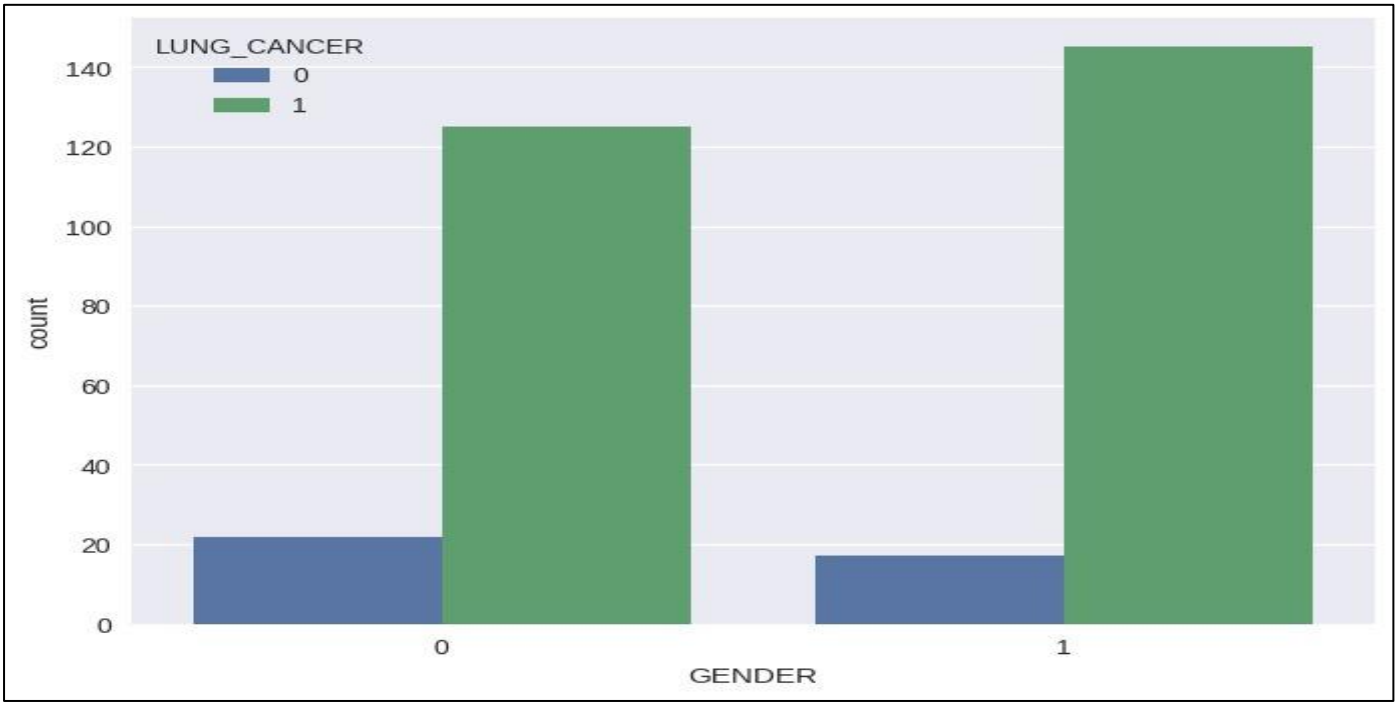

Fig 3 Analysis of Gender vs Lung Cancer

To properly assess model performance the dataset was divided into training and testing sets following preprocessing and feature selection. Using the conventional 80:20 train-test split ratio, 80% of the data was utilized to train machine learning algorithms and 20% was set aside for testing. Cross-validation methods like k-fold cross-validation (with k=5 and k=10) were also employed to make sure the model was assessed on several data subsets which decreased overfitting and enhanced generalization. To guarantee that both classes lung cancer-positive and lung cancer-negative cases were fairly represented in the training and testing sets the stratified split approach was used. An 80:20 split was used to further separate the training dataset into training and validation sets for deep learning models. Using early stopping and dropout approaches the validation set was used to track the model's learning progress and avoid overfitting. In order to improve the deep learning model's robustness and guarantee that it generalizes well to new data, data augmentation techniques like Gaussian noise addition and small perturbations were used. To find the best model for predicting lung cancer a variety of machine learning techniques, such as Decision Trees, Random Forest and Deep Neural Networks were trained using the pre-processed and appropriately divided dataset.

IV. PROPOSED METHODOLOGY

Starting with gathering data, preprocessing, feature engineering and model selection the suggested technique for lung cancer prediction proceeds in a structured pipeline before evaluation and implementation. First, a dataset was collected that included demographic information (gender, age), lifestyle factors (frequency, duration of smoking), symptoms (chest pain, coughing) and blood biomarkers (white blood cell count, hemoglobin levels). To find missing values, identify outliers, and comprehend the distribution of features, exploratory data analysis or EDA was used. Label encoding and one-hot encoding were used to encode categorical variables and suitable imputation techniques were used to handle missing values. Feature selection techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were employed to remove redundant features and normalization techniques like Min-Max Scaling and Standard

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

Scaling were utilized to guarantee model performance. Stratified sampling was then used to ensure balanced class representation after the dataset was divided into sets for training and testing in an 80:20 ratio. Three machine learning models were used to predict lung cancer once the data had been preprocessed: Random Forest, Decision Trees and Deep Neural Networks (DNN). While the Deep Neural Network identified intricate patterns in the data, the Decision Tree and Random Forest models offered solid and understandable classification.

➢ *Deep Neural Networks*

A form of artificial neural networks called Deep Neural Networks (DNN) has several hidden layers [14] and is intended to identify intricate patterns and connections in data. DNNs can simulate highly non-linear interactions which makes them ideal for applications like image recognition, natural language processing and medical diagnosis. This is in contrast to shallow networks which are limited in their ability to capture complex dependencies. Each neuron in a DNN takes weighted inputs applies an activation function, and then sends the result to the following layer. A DNN is composed of an input layer, several hidden layers and an output layer. By iteratively modifying the model's weights, the backpropagation process and an optimization method like stochastic gradient descent (SGD) or the Adam optimizer are used to reduce the error. A deep neural network was built to predict lung cancer. It had an output layer that used a sigmoid activation function for binary classification (cancerous vs. non-cancerous instances), numerous fully connected hidden layers and an input layer that matched the number of characteristics in the dataset.

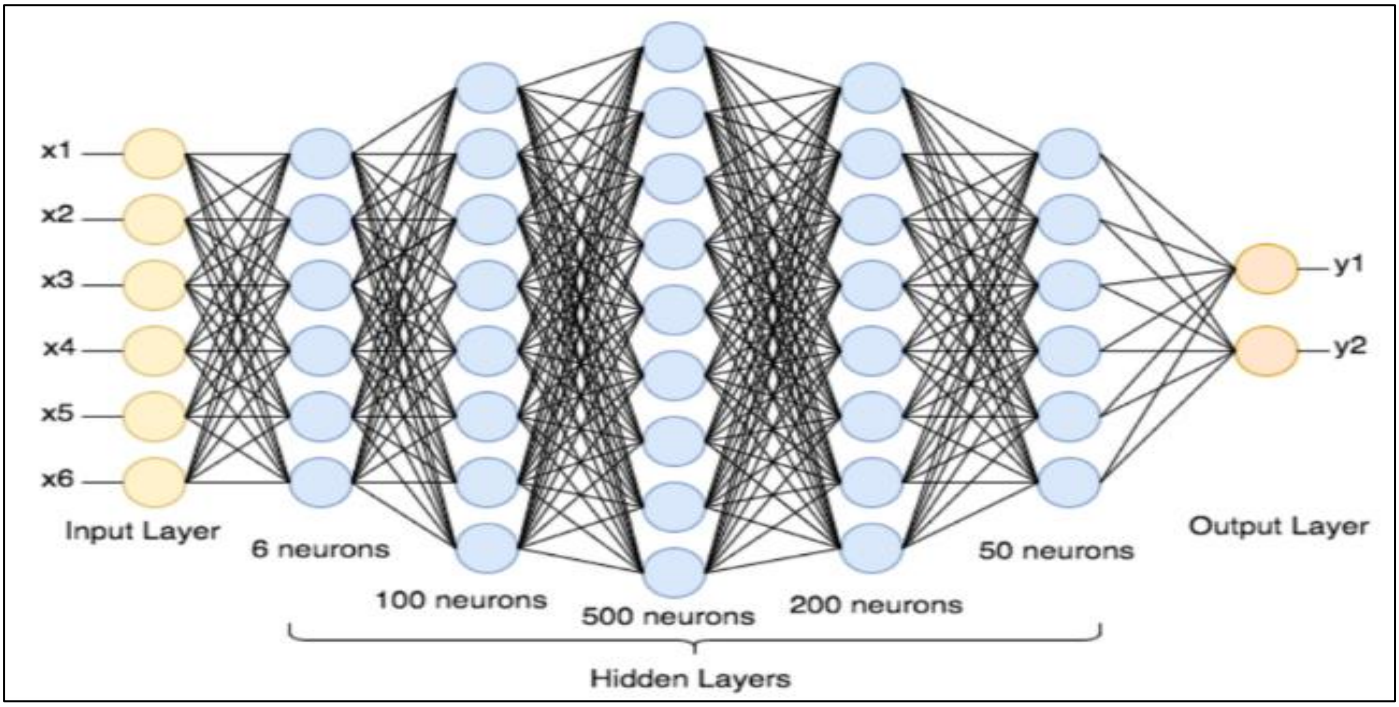

Fig 4 Deep Neural Network Architecture

An 80:20 train-test split was used to train the DNN model and an extra validation set was included to track learning and avoid overfitting. To avoid needless calculations and overfitting, early stopping was used to end training as soon as validation loss stopped improving and the Adam optimizer was selected for effective weight updates. To maximize the number of layers, neurons per layer, learning rate and dropout rate, hyperparameter tuning was done. To enhance generalization and lessen the complexity of the model regularization strategies like L2 regularization were also used. The model regularly modified its weights during the several epochs of training in order to reduce loss and increase accuracy. When compared to more conventional machine learning models like Random Forest and Decision Trees the Deep Neural Network originally showed low accuracy (12.62%) despite its capacity to represent intricate relationships in data. Suboptimal hyperparameters, imbalanced classes, or a lack of training data could all be contributing contributors to this low accuracy. In contrast to Random Forest and Decision Trees which excel at handling structured tabular data, DNNs usually need sizable datasets with significant feature representations in order to reach high accuracy. To increase predictive performance for lung cancer detection, future developments might use ensemble techniques, transfer learning or hybrid models that combine DNN with feature-based classifiers.

➢ *Random Forest*

In order to increase accuracy and decrease overfitting, Random Forest an ensemble learning technique, constructs several decision trees and aggregates their predictions [15]. It is predicated on the idea of bagging, or bootstrap aggregating in which a number of randomly chosen dataset subsets are used to train decision trees separately. Each tree contributes its output during the prediction process, and the final forecast is decided by average (for regression) or majority vote (for classification). This method lowers variance and keeps individual trees from overfitting to the training set, which aids

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

Random Forest in achieving high accuracy and robustness. A group of decision trees trained on various dataset subsets were used to build the Random Forest model for lung cancer prediction. To guarantee variation among trees characteristics were picked at random at each node split and each tree was trained using a randomly chosen subset of the training data. Complex correlations between demographic variables (age, gender), lifestyle choices (history of smoking), symptoms (chest pain, coughing) and blood report parameters (white blood cell count, hemoglobin) were all well captured by the model. By averaging the outputs of several decision trees, Random Forest generalizes well and improves performance and stability in contrast to single decision trees, which are prone to overfitting.

Accuracy, mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to assess the Random Forest model after it was trained on an 80:20 train-test split. With a high accuracy of 93.54%, it outperformed the Deep Neural Network (12.62%) and Decision Tree (90.32%) by a wide margin. With an MSE of 0.06451, MAE of 0.06451 and MAPE of 0.056451, the model's error metrics were likewise low, suggesting accurate and trustworthy predictions. Random Forest's success can be ascribed to its capacity to effectively manage numerical and categorical data while reducing the possibility of overfitting via ensemble learning. A Bagging Classifier was used to improve Random Forest's performance even further. By training several Random Forest instances on various bootstrapped datasets and averaging their results bagging assisted in lowering variance. By enhancing model stability and resilience, this method made sure that slight variations in the dataset had no impact on prediction accuracy. Random Forest's predictive power was further enhanced by the Bagging Classifier which made it an extremely successful model for the identification of lung cancer. Random Forest is a useful tool for medical applications because of its high accuracy and excellent interpretability which help with early detection and enhance patient outcomes.

- *Decision Tree*

A popular supervised learning technique, Decision Tree uses input information to create judgments in a tree-like structure. Using criteria like the Gini Index or Information Gain (Entropy) it iteratively divides the dataset into subsets according to the most important attribute. The model's nodes stand in for features, its branches for decisions and its leaf nodes for the final classification or forecast. Decision trees are useful in medical diagnostics when decision-making transparency is essential because they are easy to comprehend and analyze. They are susceptible to overfitting, too, particularly if the tree is overly deep and picks up noise from the data rather than significant patterns. Using patient demographics, smoking patterns, symptoms and blood test values as input features a Decision Tree classifier was constructed for the purpose of predicting lung cancer. In order to maximize the separation between lung cancer-positive and negative instances, the algorithm identified the optimal splits based on Gini Impurity. By using a sequence of "yes" or "no" choices, the tree structure allowed for effective categorization and produced an interpretable model. The algorithm successfully distinguished between malignant and non-cancerous cases by identifying important risk variables like smoking history, chronic coughing and abnormal blood levels.

The Decision Tree model achieved an accuracy of 90.32%, which was lower than Random Forest (93.54%) but higher than the Deep Neural Network (12.62%). It was trained using an 80:20 train-test split and assessed using a variety of performance indicators. With Mean Squared Error (MSE) of 0.09677, Mean Absolute Error (MAE) of 0.09677 and Mean Absolute Percentage Error (MAPE) of 0.0725 the model's error levels were comparatively low. Even while the model worked well it tended to overfit the training set which made it less generalizable when applied to fresh patient data. To restrict the tree's depth and improve its generalizability, pruning strategies including cost complexity pruning were used. Ensemble approaches such as Random Forest which mix several Decision Trees to increase accuracy and stability were investigated in order to solve the overfitting problem and strengthen the prediction's resilience. Optimizing model performance also involved adjusting hyperparameters such maximum depth, minimal samples per split and minimum data per leaf. In order to eliminate redundant attributes and increase prediction efficiency future improvements might incorporate feature selection approaches. The Decision Tree is still a useful tool in medical applications because of its transparency and interpretability which help medical professionals make well-informed decisions regarding the diagnosis of lung cancer.

V. RESULTS

Important criteria like accuracy, mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to assess the performance of the lung cancer prediction models. In terms of accuracy and error reduction Random Forest performed better than the Decision Tree and Deep Neural Network (DNN) models among those that were examined. The DNN model struggled with an accuracy of only 12.62% while the Random Forest classifier scored the best accuracy of 93.54% followed by the Decision Tree at 90.32%. While traditional machine learning models perform well in feature-based classification tasks the DNN model's poor performance indicates that deep learning might not be the best strategy for this structured dataset. Error measurements offer more in-depth information about the models' dependability. The Random Forest model produced fewer mistakes than the others as indicated by the MSE values of 0.06451 for Random Forest, 0.09677 for Decision Tree and 0.8747 for DNN. The same pattern was seen in the MAE values, which were 0.8743 for DNN, 0.06451 for Random Forest and 0.09677 for Decision Tree. The deep learning method was less successful for this dataset as evidenced by the MAPE values which were greatest for DNN (43.7177), lowest for Random Forest (0.056451) and second for Decision Tree (0.0725). These findings show that, in comparison to standalone models ensemble approaches such as Random Forest offer more reliability for predicting lung cancer.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

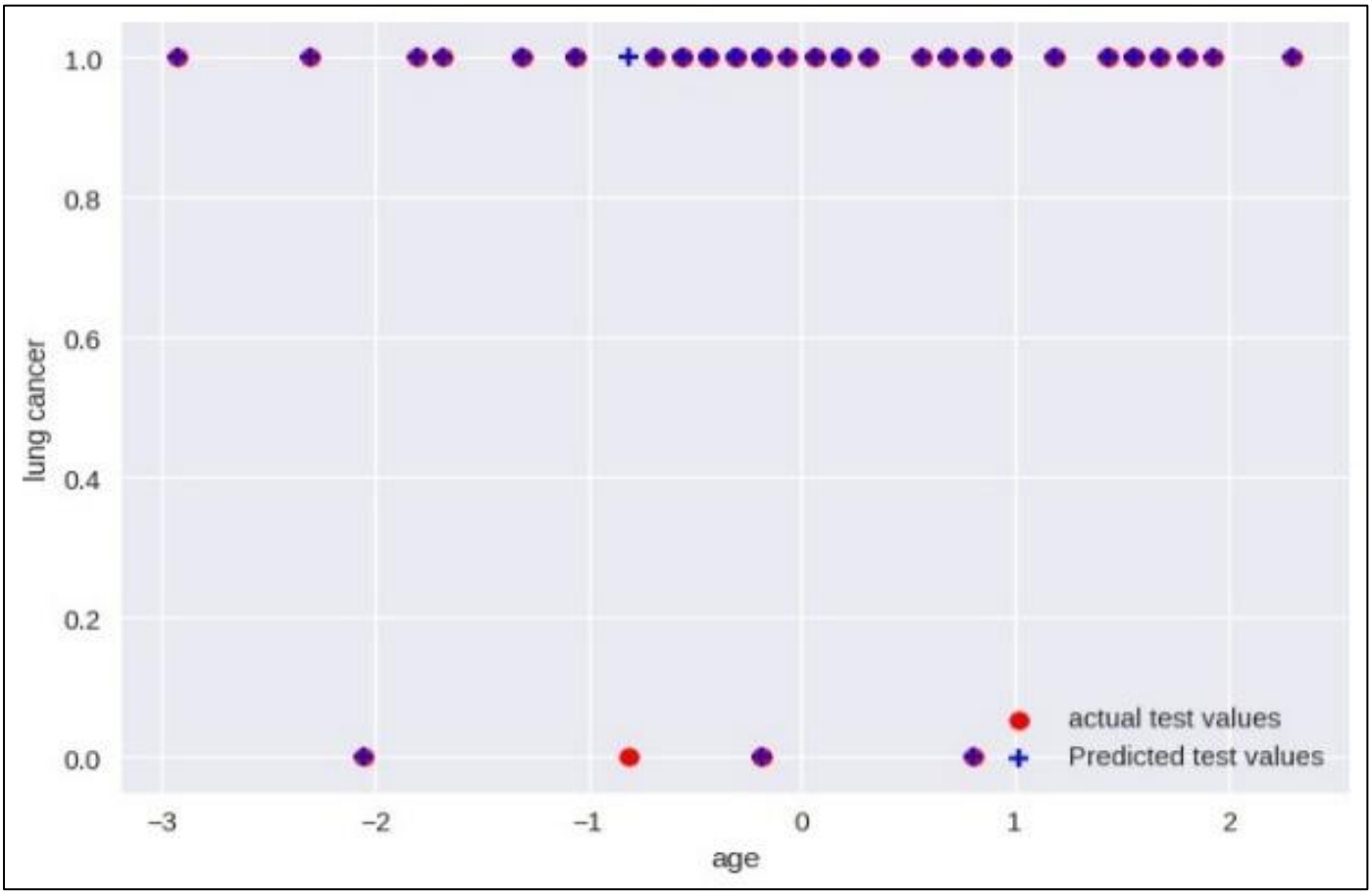

Fig 5 True vs False Predictions of Random Forest

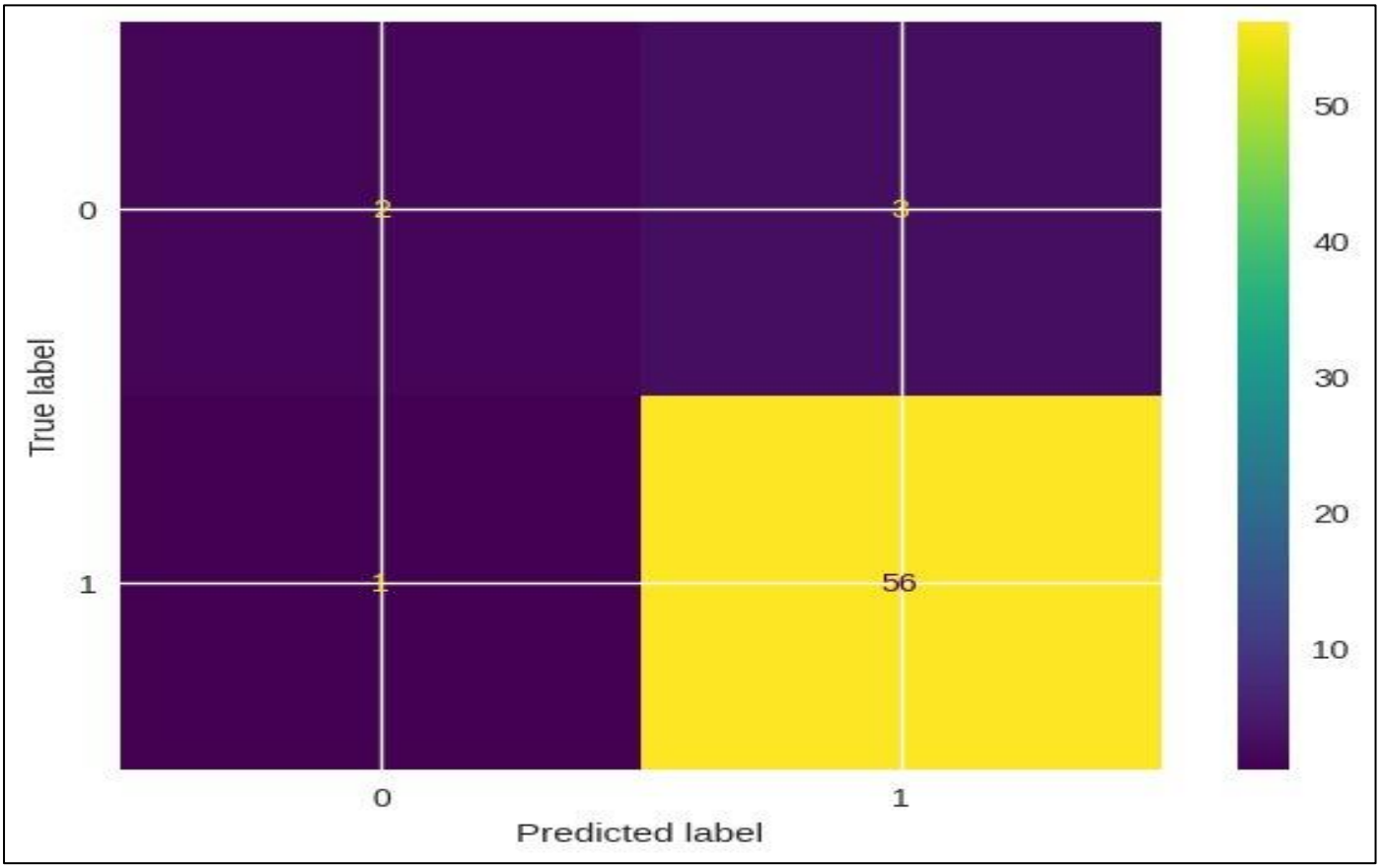

Fig 6 Confusion Matrix of Random Forest Classifier

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

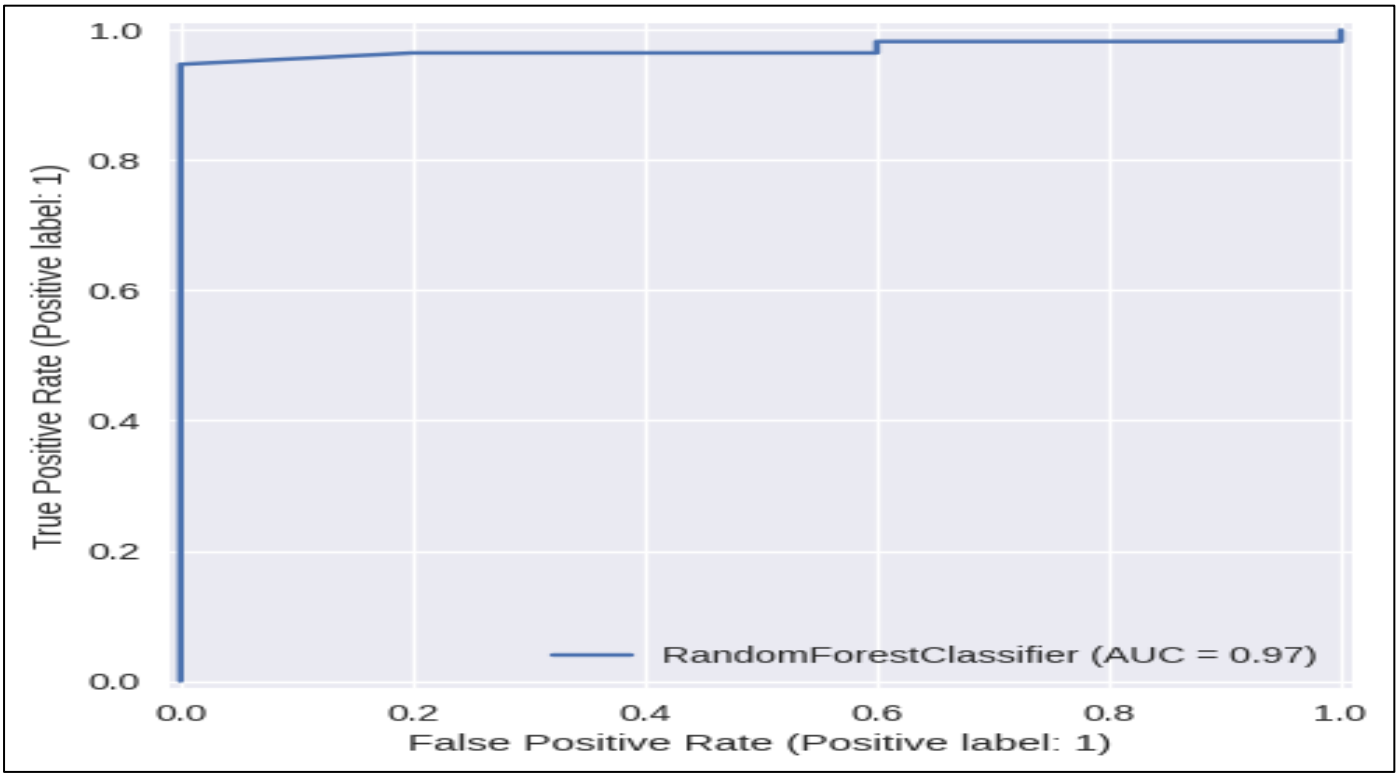

Fig 7 ROC Curve of Random Forest Classifier

Different plots and visualizations were created to show the variations in model performance. A bar chart illustrating the accuracy comparison demonstrated Random Forest's higher performance. The trend of MSE, MAE and MAPE for the various models was displayed using line graphs which revealed that DNN had noticeably larger mistakes than Random Forest which had the lowest error rates. To aid in the analysis of model misclassifications a confusion matrix displaying true positive, true negative, false positive and false negative values was also produced for every model. The capacity of decision tree-based models to offer feature importance rankings is one of their benefits. To determine the most important characteristics in the prediction of lung cancer, a feature significance plot was made for the Random Forest model. According to the findings the main factors influencing the identification of lung cancer were age, smoking behaviors, hemoglobin levels, white blood cell count and chronic coughing. Because it highlights important risk variables that should be taken into account when diagnosing patients this knowledge is essential for medical practitioners.

Fig 8 User Interface

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

Streamlit was used to create an intuitive user interface that allowed for real-time lung cancer prediction. Users can enter patient information such as age, gender, smoking history, symptoms and blood report values using the user interface (UI), and the trained models will predict the outcome. Real-time findings are displayed on the interactive interface, indicating the patient's chance of developing lung cancer. Users can also see confidence levels and probability scores, which give them an idea of how confident the model is in its forecast. To improve the user experience, the Streamlit UI has a number of interactive elements. To provide simple data entry, input values for numerical and categorical categories were captured using sliders and dropdown menus. The model prediction is triggered by a submit button, and the results are shown immediately. Users can compare various models thanks to the graphical plots that display the model performance in the user interface. Additionally, a feature importance section helps users better evaluate the data by highlighting the important elements influencing lung cancer forecasts.

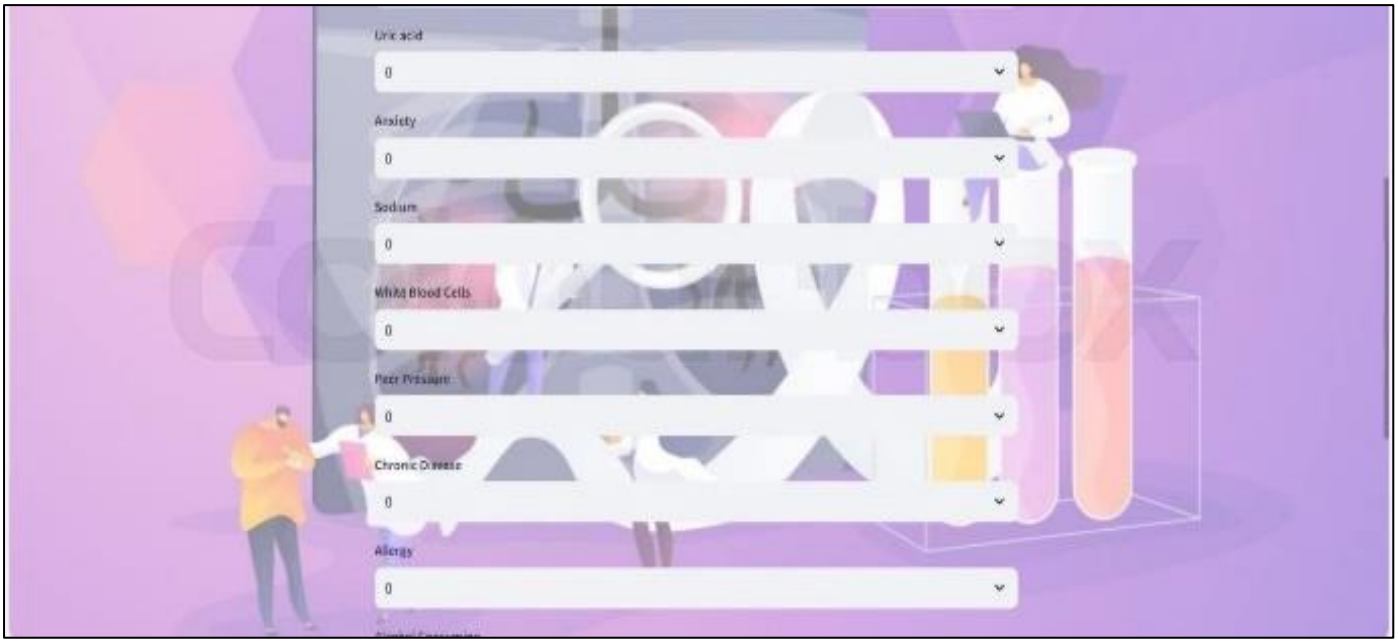

Fig 9 Inputs for Lung Cancer Prediction

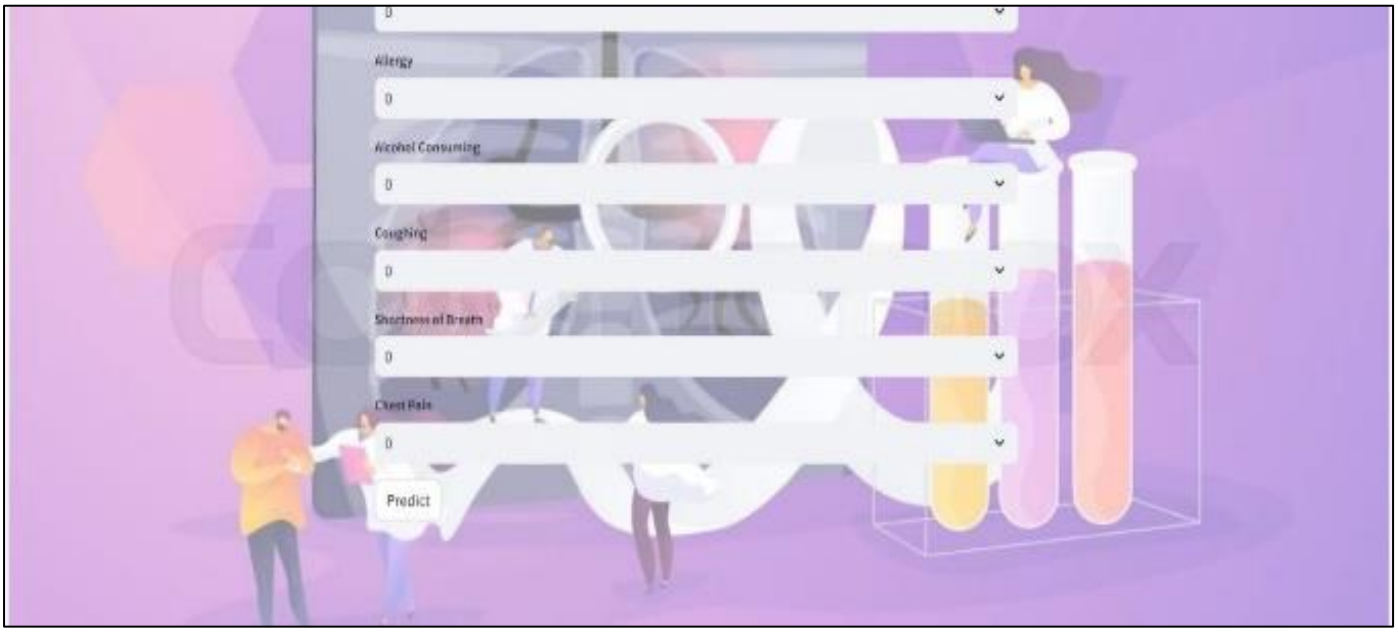

Fig 10 Analysis Lung Cancer Prediction

Model selection, where users can select between Decision Tree, Random Forest, or DNN for prediction is one of the main features built into the user interface. This enables users to evaluate various models and decide which one best suits their needs. In order to ensure that people are aware of the prediction's dependability the user interface additionally shows model accuracy, error metrics and a confusion matrix. Users may see why Random Forest is the top-performing

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun173

model by comparing accuracy and error rates visually. The experimental findings demonstrate that for structured medical datasets, ensemble learning techniques such as Random Forest perform noticeably better than stand-alone models like Decision Trees and Deep Neural Networks. Due to the feature-based structure of this tabular dataset DNN models did not perform well despite their strength in picture and text-based data. Both patients and medical professionals can use the Streamlit UI's user-friendly platform for real-time lung cancer risk prediction. To improve model robustness and generalizability future developments can include adding more patient records to the dataset combining deep learning hybrid models, or integrating sophisticated ensemble approaches.

VI. CONCLUSION

In this study, we used patient demographics, symptoms, and blood test results to create a machine learning-based lung cancer prediction system. The most dependable model for classifying lung cancer was Random Forest which obtained the highest accuracy (93.54%) with the lowest error rates when compared to Deep Neural Networks (DNN), Decision Trees and Random Forest. Because the dataset was organized, the DNN model had trouble but the Decision Tree model did well as well. For real-time prediction an intuitive Streamlit interface was used enabling users to enter patient information and get immediate results. In this study we developed a machine learning-based lung cancer prediction system based on patient demographics, symptoms and blood test data. When compared to Deep Neural Networks (DNN), Decision Trees and Random Forest, Random Forest was the most reliable model for lung cancer classification with the highest accuracy (93.54%) with the lowest error rates. The DNN model struggled because of the ordered dataset however the Decision Tree model also performed well. An easy-to-use Streamlit interface was employed for real-time prediction, allowing users to input patient data and receive results instantly.

REFERENCES

[1]. Camell, Christina D., et al. "Senolytics reduce coronavirus-related mortality in old mice." Science 373.6552 (2021): eabe4832.
[2]. Ahmed, Naveed, et al. "Smoking a dangerous addiction: a systematic review on an underrated risk factor for oral diseases." International Journal of Environmental Research and Public Health 18.21 (2021): 11003.
[3]. Tárnoki, Dávid László, et al. "Lung imaging methods: indications, strengths and limitations." Breathe 20.3 (2024).
[4]. Bhuiyan, Mohammad Shafiquzzaman, et al. "Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models." Journal of Computer Science and Technology Studies 6.1 (2024): 113-121.
[5]. Ahmed, Usman, et al. "Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems." Scientific Reports 14.1 (2024): 30532.
[6]. Gong, Yichen, et al. "Figstep: Jailbreaking large vision-language models via typographic visual prompts." arXiv preprint arXiv:2311.05608 (2023).
[7]. Folpe, Andrew, and G. Petur Nielsen, eds. Bone and Soft Tissue Pathology E-Book: A Volume in the Foundations in Diagnostic Pathology Series. Elsevier Health Sciences, 2022.
[8]. Boddapati, Mohan Sai Dinesh, et al. "Creating a protected virtual learning space: a comprehensive strategy for security and user experience in online education." International Conference on Cognitive Computing and Cyber Physical Systems. Cham: Springer Nature Switzerland, 2023.
[9]. Dinesh, Paidipati, A. S. Vickram, and P. Kalyanasundaram. "Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy." AIP Conference Proceedings. Vol. 2853. No. 1. AIP Publishing, 2024.
[10]. West, Tim, et al. "A blood-based diagnostic test incorporating plasma Aβ42/40 ratio, ApoE proteotype, and age accurately identifies brain amyloid status: findings from a multi cohort validity analysis." Molecular neurodegeneration 16.1 (2021): 30.
[11]. Plichta, Jennifer K., et al. "Implications of missing data on reported breast cancer mortality." Breast cancer research and treatment 197.1 (2023): 177-187.
[12]. Sinsomboonthong, Saichon. "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification." International Journal of Mathematics and Mathematical Sciences 2022.1 (2022): 3584406.
[13]. Kurita, Takio. "Principal component analysis (PCA)." Computer vision: a reference guide. Cham: Springer International Publishing, 2021. 1013-1016.
[14]. Li, Zhuo, Hengyi Li, and Lin Meng. "Model compression for deep neural networks: A survey." Computers 12.3 (2023): 60.
[15]. Palimkar, Prajyot, Rabindra Nath Shaw, and Ankush Ghosh. "Machine learning technique to prognosis diabetes disease: Random forest classifier approach." Advanced computing and intelligent technologies: proceedings of ICACIT 2021. Singapore: Springer Singapore, 2021. 219-244.