International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

# Predicting and Classifying Lung Cancer Using KNN Classification Machine Learning Algorithm

Oluwagbemiga O. Shoewu[1], Adetokunbo Adenowo[2], Oluwafemi J. Ayangbekun[3], Adeola A Adelakun[4], Abd-Rahman A. Badmus[5]

[1,2,5]*Department of Electronics and Computer Engineering, Lagos State University, Epe, Lagos State, Nigeria.*
[3]*Department of Computer Engineering, Federal polytechnic Ayede, Oyo State, Nigeria*
[4]*Department of Electrical Electronics Engineering,Ogun State Institute of Technology, Igbesa, Ogun State, Nigeria*

---

Date of Submission: 25-08-2025 Date of Acceptance: 05-09-2025

---

**ABSTRACT**

Lung cancer remains one of the most prevalent and deadly forms of cancer globally, necessitating early detection and accurate prediction to improve patient outcomes. This paper presents the development of a machine learning-based model to predict lung cancer levels—low, medium, or high—using a dataset containing various clinical and lifestyle attributes such as age, gender, air pollution exposure, smoking habits, and genetic risk. Multiple machine learning algorithms, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and others, were applied to the dataset to assess their prediction performance. The results revealed that Random Forest (RF) yielded the highest accuracy of 100%, precision of 98%, sensitivity of 97%, and an F1-score of 95%, making it the most effective model for lung cancer prediction. Decision Trees (DT) followed with an accuracy of 88%, while K-Nearest Neighbors (KNN) and Logistic Regression (LR) achieved accuracies of 87% and 81%, respectively. Kernel Density Estimate (KDE) plots were utilized to assess the distribution of features, providing insights into feature selection and class separation. The study highlights the significance of accurate feature selection and the efficacy of machine learning models in predicting lung cancer severity, paving the way for better diagnostic tools in medical practice.

**Keywords**: Algorithm, lung Cancer, Machine Learning, Model

## I. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with early diagnosis playing a crucial role in improving patient outcomes and survival rates [1]. Traditional diagnostic methods, although effective, often involve invasive procedures and extensive medical examinations that can delay the identification and treatment of the disease [2]. In this context, machine learning (ML) presents a promising avenue for enhancing diagnostic accuracy and efficiency [3]. This paper focuses on utilizing machine learning algorithms including the K-Nearest Neighbors (KNN), Logistic Regression, and other classification algorithm, which are widely used and robust ML technique [4], to classify and predict lung cancer [5]. The KNN algorithm operates on the principle of identifying the nearest data points in a multidimensional space to classify a given instance based on the majority class among its neighbors. Its simplicity, coupled with its ability to adapt to different types of data distributions, makes KNN an ideal candidate for medical diagnosis applications.

The primary objective of this study is to develop a machine learning classification model capable of accurately distinguishing between malignant and benign lung cancer cases. By leveraging a comprehensive dataset of patient records and diagnostic imaging features, the model aims to learn and identify critical patterns associated with lung cancer. This approach has the potential to significantly reduce the time and resources required for accurate diagnosis, thereby facilitating earlier intervention and treatment. This paper explores various aspects of the machine learning models, including optimal parameter selection, feature importance analysis, and performance evaluation metrics such as accuracy, precision, recall, and F1 score. Thereafter, the model performance will be compared with other ML algorithms to validate its effectiveness and robustness in predicting lung cancer.

The input dataset comprises attributes such as Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupation, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoking, Chest Pain, Coughing, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Fingers, Frequent Cold, Dry Cough, and Snoring. The outcome of the lung cancer prediction is categorized into three distinct levels: low, medium, and high, reflecting the varying degrees of risk associated with the disease as depicted in Fig. 1.

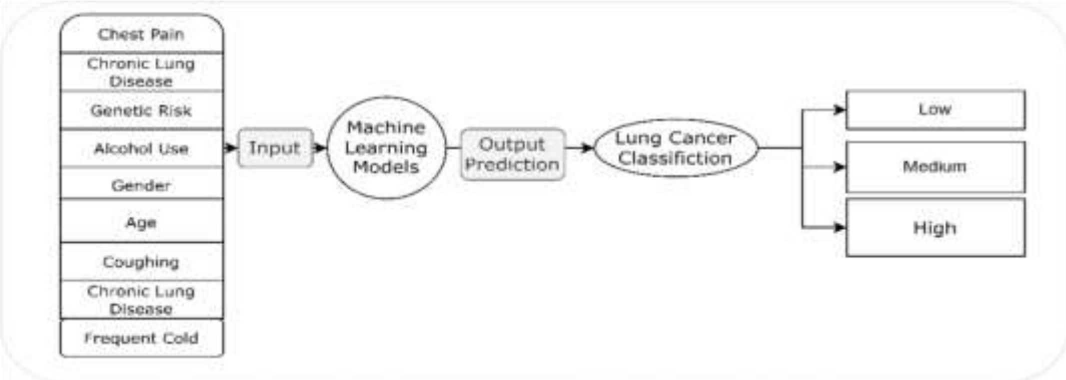



**Figure 1: Input into the machine learning algorithms.**

## II. LITERATURE REVIEW

### 2.1 BACKGROUND

The classification and prediction of lung cancer have undergone remarkable transformations over the past few decades, largely driven by advancements in machine learning (ML) and data science [6]. Lung cancer remains one of the most prevalent and lethal forms of cancer worldwide, presenting significant challenges in early diagnosis and accurate prediction [7]. The disease's complex nature, characterized by its diverse subtypes and varying biological behaviours, complicates both clinical assessment and therapeutic decision-making [8]. The utilization of machine learning algorithms in healthcare significantly enhances disease diagnosis and management by enabling accurate predictions based on patient data. These methods facilitate early detection of conditions like lung cancer, support timely interventions through risk assessment, and improve overall diagnostic accuracy by analysing complex patterns in diverse patient information. As a result, they contribute to more personalized treatment plans and ultimately lead to better patient outcomes.

By early 2010s, application of Machine Learning (ML) algorithms to lung cancer classification and predictions began to show promising results; with researchers utilizing machine learning algorithms to analyse various types of medical data, including imaging data, genetic profiles, and patient demographics. This early applications of ML to lung cancer focused on developing algorithms that could assist radiologists by highlighting suspiciousareas in medical images. These initial efforts laid the groundwork for more complex predictive models; as ML algorithms models aimed to classify lung cancer subtypes and predict disease progression (Figure 2) with greater accuracy than traditional methods.

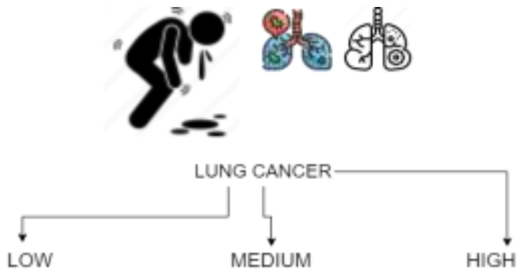



Figure 2: Lung Cancer ML prediction model.

### 2.2 Healthcare Medical Imaging Predictive Analysis

This is one of the notable applications of machine learning in healthcare [9]; as machine learning has a profound impact on the healthcare sector, unlocking new opportunities for medical research, diagnosis, and treatment methodologies. Advanced machine learning algorithms have been specifically trained to analyse and interpret medical images, such as X-rays, MRIs, and CT scans. These algorithms can detect anomalies and irregularities that may escape the attention of human practitioners, potentially leading to earlier diagnosis and treatment of various conditions. By automating image analysis, healthcare professionals can improve diagnostic accuracy and reduce the burden of manual interpretation [10].

Meanwhile, in terms of PredictiveAnalytics**,** utilization of machine learning algorithms by healthcare providers can forecast the likelihood of specific health conditions or diseases in patients by analysing their medical histories, demographic data, and other relevant factors. This predictive capability enables early intervention and personalized care, ultimately improving patient outcomes. For example, predictive models can be employed to assess the risk of hospital readmission, the onset of chronic diseases, and the effectiveness of treatment plans [11].

In summary, machine learning holds transformative potential for the healthcare industry by improving the accuracy of diagnosis, treatment efficacy, and patient outcomes. However, it is vital to recognize that the performance of these algorithms is heavily contingent upon the quality and integrity of the data used for training. Moreover, ethical considerations must be addressed when implementing machine learning in healthcare to ensure that the technology is applied responsibly and equitably, safeguarding patient welfare and privacy [12]. The Key machine learning algorithms that are commonly employed in various applications include: Logistic Regression[13], LinearRegression[14], RandomForest[15], SupportVectorMachines **(**SVMs**)** [16], and DecisionTrees [17]. Finally, the research described in [18] highlighted the effectiveness of machine learning algorithms in medical / healthcare predictions; as the main objective of this paperis to identify diagnostic approaches by analysing clinical data and employing classification analysis alongside machine learning methods to facilitate early chronic disease detection. Meanwhile, both machine learning and deep learning emphasizes equipping computers with the ability to learn from data inputs and improve their performance over time [19].

## III. METHODOLOGY

### 3.1 Methodological Overview

**Applied structural approach is to ensure the effective development and implementation of the models for lung cancer prediction as depicted in Figure 3. This includes: Datasets Collection, data pre-processing, data splitting, model selection, model training, model Evaluation and Implementation**

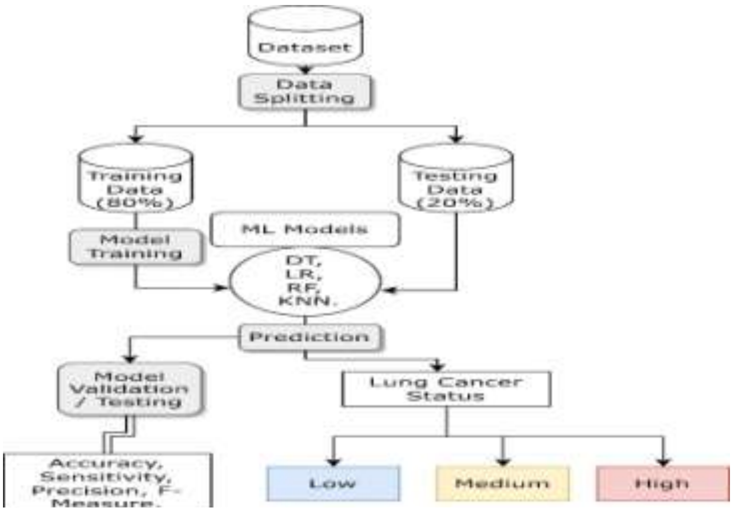



**Figure 3: Overview of this research**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

### 3.2 Simulation sequential Phase

The machine learning modelling is being generated for ML classifiers to predict based on the multi class target variable is as shown below (Figure 4).The head of the dataset shows the first 5 rows or entries. However, it should be noted that a machine learning model can only work with numerical values which gives rise to the necessity to replace the multi-target class variables of Low, Medium and High with 1, 2 and 3 as shown in Fig 4.

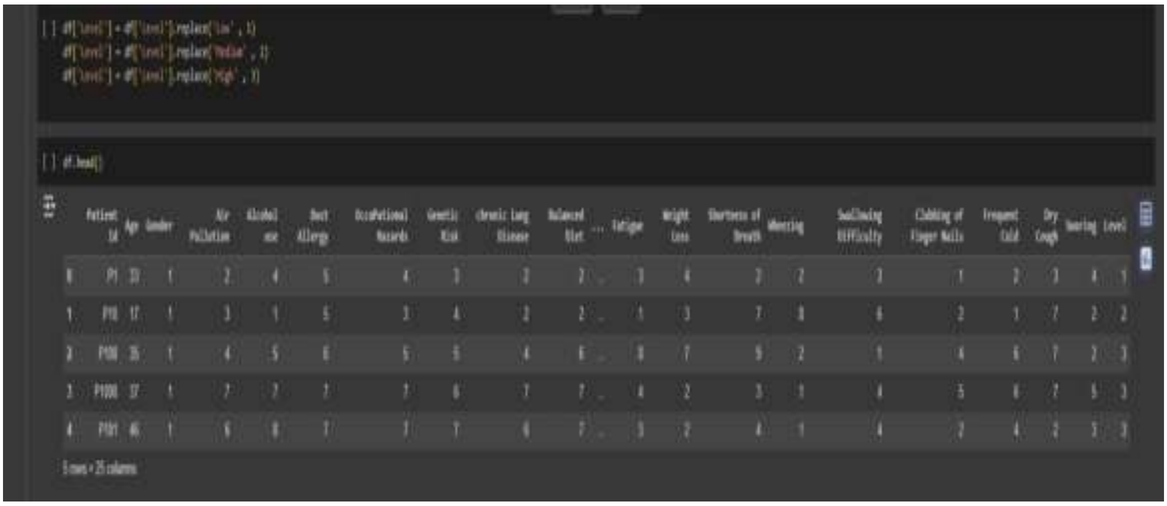


**Figure4: Dataset head with replaced target variables.**

With the training percentage of 0.8 and a test percentage of 0.2 of the data; the X train data were used to train the machine model while the X test data was used in the validation and testing phase of the machine learning process. These train and test data were defined to learn, establish patterns during the training phase and evaluate the model's performance by comparing the predicted values against these actual values (Figure 5).

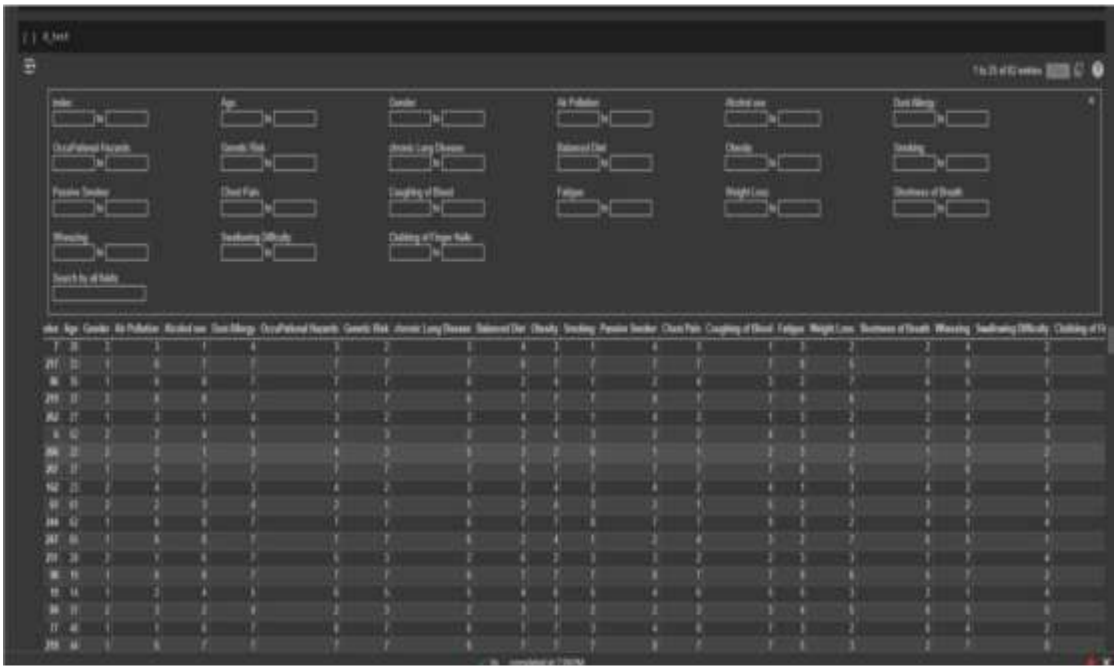


**Figure 5: X test Segment of the data**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

## IV. RESULTS AND DISCUSSION

Using machine learning algorithms for the prediction of lung cancer disease in terms of visualisation of evaluation Metrics scores of Classifiers, the following numerical results were obtained for the adopted machine learning paradigms/tools.

### 4.1 Confusion Matrices

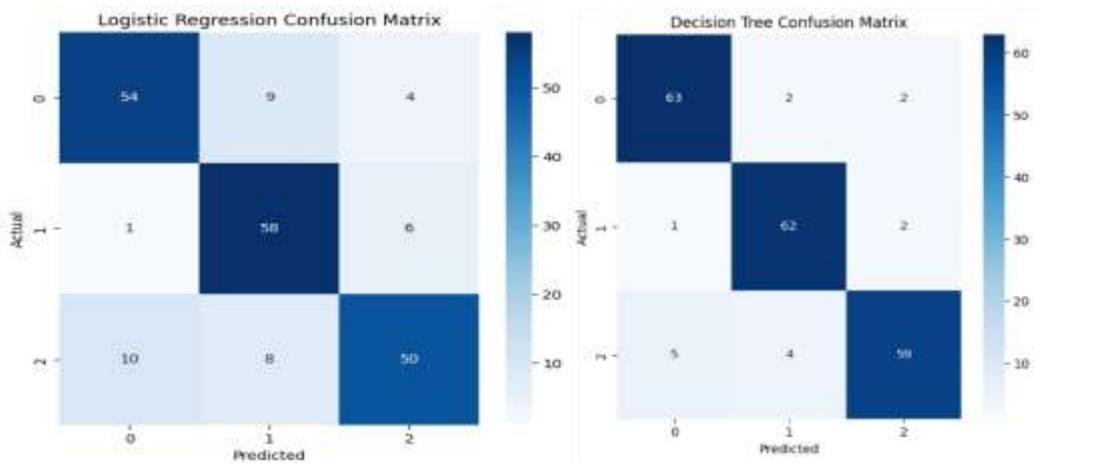
The results from the confusion matrices provide valuable insights into the classification performance of each algorithm on the lung cancer dataset, as shown in Figure 6. The results indicate comparison of the outcomes for Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN).

The Logistic Regression (LR) model shows moderate performance. While it correctly classifies the majority of instances in each class, there is noticeable confusion between classes, particularly between class 1 and class 2, where 9 instances of class 1 were misclassified as class 2, and 6 instances of class 2 were misclassified as class 3 (Figure 6). This suggests that LR struggles to differentiate between certain categories, especially in the case of adjacent or overlapping feature sets. However, it still achieves a relatively high number of correct classifications for each class.

Decision Tree (DT) model demonstrates better classification accuracy than LR, as evidenced by fewer misclassifications. The confusion between classes is significantly reduced, particularly for class 1 and class 2, where only 2 and 1 instances were misclassified, respectively (Figure 6). The clear separation of classes suggests that DT is able to handle the complexity of the feature space better than LR. Additionally, the higher accuracy across all classes implies that DT can capture nonlinear relationships in the data more effectively.

However, from Figure6, the Random Forest (RF) model exhibits near-perfect classification, with no misclassifications observed in any class. This exceptional performance is expected, as RF is known for its robust ability to generalize by combining multiple decision trees. RF's ability to handle variance and avoid over-fitting results in high predictive accuracy and clear differentiation among the classes. The absence of errors in all categories shows that RF is the best-performing algorithm for this particular dataset, achieving superior generalization across the classes.

Finally, from Figure 6, the K-Nearest Neighbors (KNN) model performs similarly to Logistic Regression, showing some confusion between classes, particularly between class 1 and class 2. KNN misclassifies 6 instances of class 1 as class 2, and 7 instances of class 3 as class 2. Although KNN performs well overall, its confusion in separating closely related classes, as seen with misclassifications in both class 1 and class 3, indicates that the distance-based nature of KNN struggles when classes are not linearly separable or when data points are close to decision boundaries.



International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

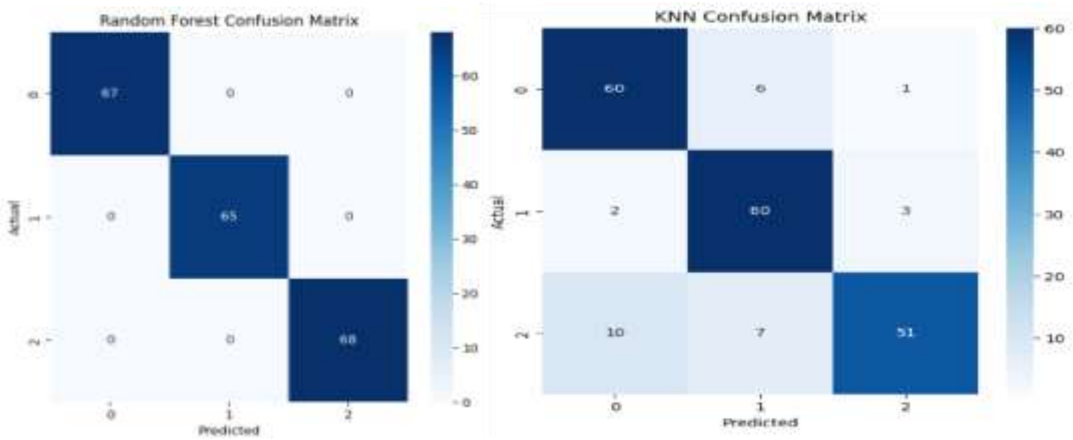



Figure 6: Predictions by Logistic Regression, Decision Tree, Random Forest and KNN.

Random Forest outperforms all other algorithms, showing no misclassifications across the board, making it the best choice for this dataset. Decision Tree follows, with very few misclassifications, while Logistic Regression and KNN exhibit more errors, particularly in distinguishing between classes 1 and 2. Also both Random Forest and Decision Tree exhibit strong separation between the classes, while Logistic Regression and KNN have more difficulty, especially in distinguishing between adjacent or similar classes (e.g., class 1 and class 2).

Meanwhile, Random Forest's performance indicates its robustness in handling both the variance and complexity of the dataset. Decision Tree also performs well but is slightly more prone to misclassification compared to Random Forest. In contrast, Logistic Regression and KNN may require more feature engineering or tuning to improve class differentiation.

### 4.2 Evaluation Metric Scores

The evaluation metrics of Accuracy, Precision, Sensitivity (recall), and F1-Score for the four machine learning models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN); reveal notable differences in their performance on the lung cancer dataset (Figure 7). The performance metrics classifiers of the resulting outcomes and relationships is also indicated using line graph representation showing comparative evaluation scores (Figure 8).

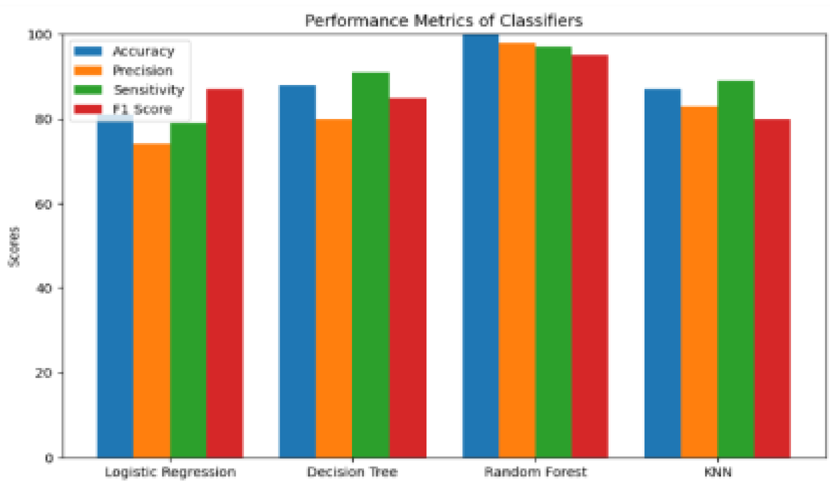



**Fig. 7: Showing comparative general evaluation scores of each classifier.**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

Logistic Regression (LR) demonstrates moderate performance, with an accuracy of 81%. Its precision, at 74%, indicates that it is relatively conservative in making positive predictions, meaning it avoids false positives to a certain extent. However, the sensitivity (79%) shows that it is reasonably good at identifying true positives, though some cases are missed. The F1-score (87%), which balances precision and recall, is relatively high, suggesting that LR maintains a good trade-off between false positives and false negatives. This model performs adequately but struggles compared to more complex algorithms when differentiating between classes.

The Decision Tree model improves upon Logistic Regression, achieving an accuracy of 88%. It performs well in terms of precision (80%), meaning it effectively avoids false positives. Its sensitivity is particularly strong at 91%, indicating that it correctly identifies most of the true positive cases. The F1-score (85%) suggests a good balance between precision and recall, slightly lower than LR due to the focus on sensitivity. DT handles class differentiation better than LR, as evident by its higher accuracy and sensitivity, and is more adept at identifying true positives, making it more suitable for cases where recall is crucial.

Random Forest outperforms all other models, with a perfect accuracy of 100%, indicating that it correctly classified all instances in the dataset. The precision (98%) and sensitivity (97%) are extremely high, showing that RF makes very few false positive and false negative errors. The F1score (95%) reflects the strong balance between precision and recall, further demonstrating that Random Forest is an ideal model for this dataset. Its ensemble approach, combining multiple decision trees, makes it highly robust and effective in capturing the complexities of the data. RF is the top performer across all metrics, highlighting its ability to generalize well and classify correctly in all cases.

K-Nearest Neighbors (KNN) achieves an accuracy of 87%, which is slightly lower than Decision Tree but still respectable. Its precision (83%) and sensitivity (89%) indicate that it balances false positives and true positives fairly well. However, its F1-score (80%) is lower than both DT and LR, suggesting that while KNN performs adequately, it is less effective than other models in handling both precision and recall simultaneously. This is likely due to KNN's reliance on proximity to classify instances, which can result in higher misclassification when there is overlap between classes or noisy data.

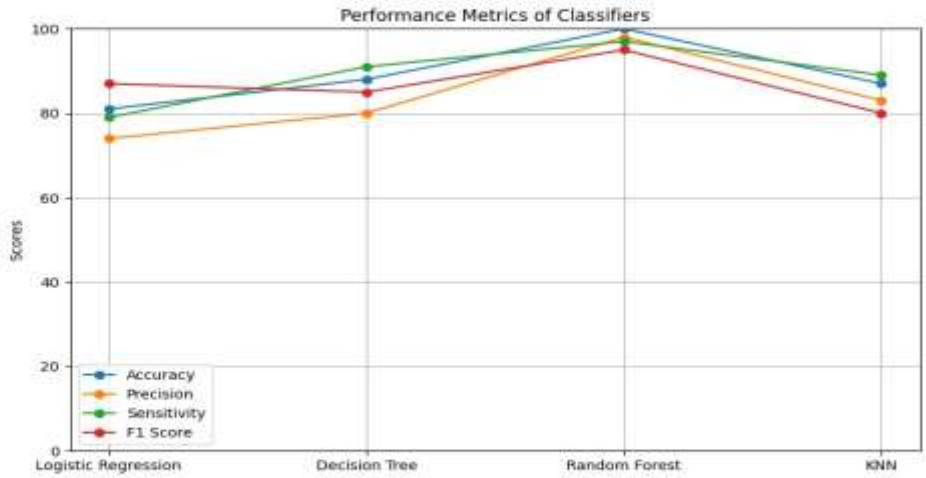



**Figure 8: Line plot showing the comparative evaluation score for each classifier.**

However, optimization of KNN's hyperparameters particularly the number of neighbors (k) and the distance metric, was crucial in enhancing model performance. The best results were thereby achieved with a moderate k value, balancing bias and variance. Finally, the model was rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC); gives high values across these metrics indicating robust performance and reliability in lung cancer classification across various classes (Figure 9). Overall, the research confirms the viability of using the KNN classification model for lung cancer prediction. The model's simplicity and effectiveness make it a valuable tool in medical diagnostics, potentially aiding early detection and improving patient outcomes.

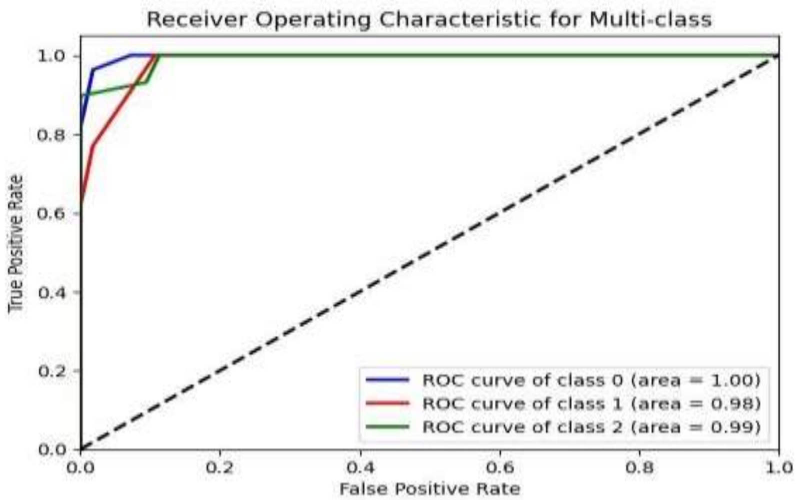



**Figure9: ROC for the predictive outcomes.**

**4.3 Kernel Density Estimation (KDE)**

Kernel Density Estimation is a probabilistic estimate distribution of random functions in terms of data visualization. It's a Probability density estimation on the available datasets using smooth density curves. In this paper, Kernel Density Estimation is being applied to lung cancer classifications and predictions for Spatial analysis in a specific region using machine learning algorithms. This is to establish and estimate the density of risk factors by causative agents like smoking prevalence, age, status etc for targeted interventionsusingKDE analysis (Figure 10) through ccareful interpretation of statistical results for clinical relevance by exploring relationships between continuous variables in lung cancer.

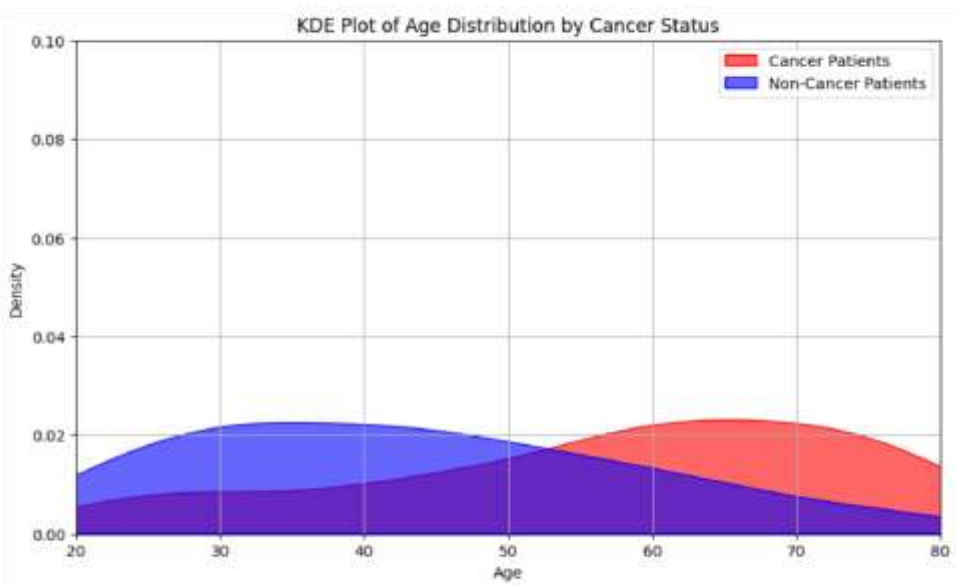



**Figure 10: KDE plot showing correlation between age and cancer status**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 933-942 www.ijaem.net ISSN: 2395-5252

## V. CONCLUSION

The research on using machine learning model to classify and predict lung cancer has provided significant insights into the applicability and performance of machine learning models / algorithm in the medical field. The primary objectives of the study were to develop a model that could accurately classify lung cancer cases based on clinical and demographic data, and to evaluate the model's performance in a real-world context. The comprehensive methodology included problem definition, data collection, pre-processing, model training, evaluation, and optimization. Random Forest clearly emerges as the best-performing model, achieving perfect accuracy and near-perfect precision, sensitivity, and F1-score. It is followed by Decision Tree, which offers a strong balance between precision and recall. KNN provides decent performance, but it struggles compared to RF and DT. Logistic Regression, while still useful, performs the weakest overall, especially in terms of precision and recall; thereby making it (LR) not be the best choice for this particular classification task. In summary, Random Forest is the best model for this dataset, followed by Decision Tree, KNN and LR.However, KNN can be a viable option depending on the application. The key findings from the research clearly shows that KNN model demonstrated a high degree of accuracy in classifying lung cancer cases in terms of model performance by leveraging the similarity between data points thereby effectively identifying patterns indicative of lung cancer. The model highlighted the significance of various features, including patient age, smoking history, genetic markers, and clinical symptoms, in predicting lung cancer; which underscores the importance of comprehensive data collection in medical diagnostics.

## REFERENCES

[1]. S. Li et al., "Deep Learning for Lung Cancer Detection and Diagnosis from Computed Tomography Images: A Systematic Review," J. Med. Syst., vol. 46, no. 10, pp. 1-13, 2022, doi: 10.1007/s10916-022-02851-6.

[2]. P. Rajpurkar et al., "Machine Learning-Based Prediction of Lung Cancer Using Clinical and Radiological Features," J. Thoracic Oncol., vol. 15, no. 5, pp. 833-842, 2020, doi: 10.1016/j.jtho.2020.04.014.

[3]. Y. Zhang et al., "Lung Cancer Prediction Using Radiomic Features and Machine Learning Algorithms," IEEE J. Biomed. Health Inform., vol. 23, no. 4, pp. 1475-1485, 2019, doi: 10.1109/JBHI.2019.2904159.

[4]. S. Hussein et al., "Deep Learning for Lung Nodule Detection and Classification in CT Images," IEEE Trans. Med. Imaging, vol. 37, no. 9, pp. 2054-2064, 2018, doi: 10.1109/TMI.2018.2833072.

[5]. Y. Li et al., "Predicting Lung Cancer Using Machine Learning and Gene Expression Data," Sci. Rep., vol. 7, no. 1, pp. 1-9, 2017, doi: 10.1038/s41598-017-04644-5.

[6]. A. Singh et al., "Lung Cancer Prediction using Machine Learning and Deep Learning Techniques," 2020 Int. Conf. Intell. Comput. Appl. (ICICA), pp. 1-6, 2020, doi: 10.1109/ICICA50836.2020.9135662.

[7]. J. Liu et al., "Deep Learning for Lung Cancer Classification and Diagnosis from CT Images," IEEE Trans. Nanobioscience, vol. 19, no. 3, pp. 361-371, 2020, doi: 10.1109/TNB.2020.2994602.

[8]. Y. Wang et al., "Machine Learning-Based Lung Cancer Prediction Using Gene Expression Data," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 17, no. 3, pp. 761-771, 2020, doi: 10.1109/TCBB.2019.2918734.

[9]. U. N. Dulhare and M. Ayesha, “Extraction of action rules for chronic kidney disease using Naïve bayes classifier,” 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Dec. 2016, doi: 10.1109/iccic.2016.7919649.

[10]. J. H. Holmes, J. S. Siewerdsen, and D. A. Jaffray, "Clinical validation of Bayesian reconstruction of a compressed-sensing (CS) MRI reconstruction algorithm for improved image quality and reduced dose," Proc. SPIE 7961, Medical Imaging 2011: Physics of Medical Imaging, Feb. 2011, doi: 10.1117/12.878247.

[11]. S. S. Ghassemi et al., "A Multitask Learning Approach for Predicting ICU Mortality Using Missing Data," IEEE J. Biomed. Health Inform., vol. 23, no. 5, pp. 1917-1926, Sept. 2019, doi: 10.1109/JBHI.2018.2870856.

[12]. N. Nassiri and E. Parimbelli, "Machine learning for healthcare: a review," ACM Computing Surveys (CSUR), vol. 53, no. 6, pp. 1-36, 2020, doi: 10.1145/3408748.
[13]. "Linear Regression," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 3234-3243.
[14]. S. Singh and S. Singh, "Machine Learning Techniques for Healthcare Data Analytics," 2021 International Conference on Innovative Computing and Communication (ICICC),
[15]. J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.
[16]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[17]. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[18]. X. Tian et al., “Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance,” Computational and Mathematical Methods in Medicine, vol. 2019, pp. 1–7, Jun. 2019, doi: 10.1155/2019/6915850.
[19]. A. K. Singh et al., "Lung Cancer Prediction using Convolutional Neural Networks and Transfer Learning," 2020 IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 1556-1563, 2020, doi: 10.1109/CVPRW50498.2020.00239.