iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62543

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Predictive Modeling for Lung Cancer Detection: Unveiling Insights through Machine Learning Techniques

Nida Aijaz Zargar[1], Anil Barnwal[2]
[1]*Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India*
[2]*Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India*

Abstract: ***Lung cancer stands as a formidable global health challenge, characterized by its widespread prevalence and high mortality rates, placing significant burdens on both public health systems and clinical practitioners. Despite notable advancements in diagnostic and therapeutic modalities, early detection remains paramount in improving patient outcomes and enhancing survival rates. In response to this imperative, the integration of predictive modelling techniques within healthcare has garnered substantial attention in recent years, offering promising avenues for bolstering early detection strategies and optimizing patient care pathways.***

The emergence of machine learning algorithms, coupled with the wealth of available healthcare data, has catalysed the development of tailored predictive models specifically designed for lung cancer prediction. This study aims to harness machine learning (ML) techniques to construct robust models capable of identifying individuals at heightened risk of developing lung cancer. By facilitating early interventions through predictive analytics, our aim is to attenuate the potential for long-term complications and ultimately enhance patient outcomes. Central to this endeavour is the generation of synthetic datasets, essential for training robust predictive models. Moreover, our methodology encompasses the evaluation of various machine learning models, selected based on rigorous performance metrics including Accuracy, Precision, Recall, and F-measure. Among the models assessed, Random Forest and Extreme Gradient Boosting emerged as the top performers, exhibiting a remarkable accuracy of 96% and an AUC of 99%. These findings underscore the efficacy of these models in predicting lung cancer risk, offering valuable insights for clinical decision-making and intervention planning.

Keywords: Lung Cancer, Random Forest, XGBoost, Accuracy, Synthetic dataset, AUC.

I. INTRODUCTION

Lung cancer represents a significant global health challenge, posing a substantial burden in terms of morbidity and mortality. Recent statistics reveal approximately 2.2 million new cases diagnosed in 2020, highlighting the critical need for advancements in detection and intervention [1]. India ranks third in cancer incidence after China and the United States, with projections indicating a 57.5% rise to 2.08 million cases by 2040 [1]. The urgency to develop innovative strategies for early detection and risk prediction is unequivocal. Machine learning, situated within the realm of artificial intelligence, has emerged as a transformative force in healthcare, presenting a profound opportunity to redefine the landscape of disease prognosis and prevention [2]. In recent years, researchers have increasingly turned to sophisticated algorithms, notably gradient boosting, to unlock the predictive power latent within vast datasets and derive meaningful insights into complex medical conditions [3] .

Traditional methods of assessing lung cancer risk have conventionally relied on demographic and clinical factors. However, the integration of machine learning techniques promises a paradigm shift towards more nuanced and accurate predictive models [3]. By leveraging a diverse array of data, encompassing genetic markers, imaging studies, and patient histories, machine learning algorithms exhibit the capacity to unveil subtle patterns and relationships that may elude traditional analyses. This capability to incorporate multifaceted information into predictive models opens new vistas for early identification of individuals at higher risk, thereby enabling timely interventions and potentially improving overall survival rates [4]. While existing staging systems, such as the TNM classification proposed by [5], provide valuable insights into disease progression, integrating machine learning into the prediction framework allows for a dynamic and data-driven approach that complements traditional methods.

©IJRASET: All Rights are Reserved

This study introduces a methodological framework aimed at crafting robust ML classification models tailored explicitly for predicting occurrences of lung cancer. The utilization of prevalent habits and symptoms/signs as pivotal input features contributes to the refinement of model efficacy.

Our principal contribution lies in a meticulous comparative evaluation of an array of classifiers, with the overarching goal of formulating an optimized model characterized by superior sensitivity and discrimination capabilities, particularly in the identification of individuals at elevated risk for lung cancer.

Throughout the rigorous evaluation of these models, a discerning selection of performance metrics was considered, encompassing precision, recall, F-Measure, accuracy, and AUC (Area Under the Curve). Additionally, the articulation and presentation of AUC ROC (Receiver Operating Characteristic) curves enrich the depth of our analytical approach. Through a meticulous examination of relevant literature and methodologies, we seek to underscore the potential impact of machine learning on the field of oncology, with a specific emphasis on its application to lung cancer prognosis and prevention This study aligns with the broader vision of leveraging technological innovations to mitigate the impact of lung cancer on global health and lays the foundation for future research endeavours in the realm of predictive medicine.

II. RELATED WORK

In a study conducted by Sobhan and Mondal [6], a pathway was proposed to identify significant lung cancer class- and patient-specific genes, utilizing two SHAP variants known as "tree explainer" and "gradient explainer." These variants were applied to the classification algorithms XGBoost and convolutional neural network (CNN) for tree-based and deep learning-based classification, respectively. The study aimed to discover population-based biomarkers, specifically class-specific top 100 genes and differentially expressed genes, shedding light on patient-specific gene sets for individuals with lung cancer. Notably, XGBoost demonstrated an impressive 96.3% accuracy in this context.

In another research endeavor [7], machine learning (ML) models were leveraged to predict the length of stay (LOS) for lung cancer patients, addressing imbalanced datasets using electronic medical records (EHR). The study utilized the MIMIC-III dataset and applied Random Forest (RF) as the primary model. The RF model, integrated with the SMOTE class balance technique, outperformed other models, emphasizing its efficacy in predicting lung cancer LOS. This approach provided insights into the significant clinical factors contributing to accurate predictions.

Jamie et al. explored the application of three eXplainable Artificial Intelligence (XAI) techniques – SHAP, LIME, and Scoped Rules – to enhance interpretability and usability of ML models on large-scale Electronic Health Record (EHR) datasets. The study utilized the synthetic dataset Simulacrum, derived from anonymized cancer data, and employed classifiers such as Logistic Regression, XGBoost, and Explainable Boosting Machine (EBM). XGBoost exhibited superior classification accuracy in this context [8].

Katarzyna et al. focused on the comparison of three lung cancer risk prediction models – BACH, PLCOm2012, and LCRAT – utilized in low-dose computed tomography lung cancer screening. The study employed selected approaches from eXplainable AI to better understand model behavior for different patients, utilizing a domestic lung cancer database [9].

Dritsas and Trigka applied various machine learning classifiers, including Naive Bayes, Random Forest, Support Vector Machine, and others, to identify individuals at high risk of developing lung cancer. The classifiers were evaluated based on accuracy, precision, recall, F-Measure, and AUC, with the Random Forest model emerging as the top performer [10].

Muntasir et al. developed ensemble learning approaches, including XGBoost, LightGBM, AdaBoost, and bagging, to forecast lung cancer. The models were validated through K-fold 10 cross-validation, and XGBoost achieved the highest accuracy of 94.42% [11].

Patra explored various machine learning classifiers for predicting lung cancer, with Radial Basis Function Network (RBF) achieving the highest accuracy of 81.25% [12].

In a study by Sim et al., health-related quality of life (HRQOL) in 5-year lung cancer survival prediction was investigated using machine learning models, including Decision Tree, Logistic Regression, Bagging, Random Forest, and AdaBoost. AdaBoost demonstrated the highest accuracy of 94.8% in this context [13].

In the study conducted by the author [14], a robust approach for the detection and classification of lung cancer utilizing CT scan images was presented. Seven classification models, including decision tree, random forest, support vector machine, naive Bayes, k-nearest neighbors, stochastic gradient descent, and multi-layer perceptron, were employed. The dataset comprised 15,750 clinical data instances, containing 6,910 benign and 8,840 malignant lung cancer-related images. The multi-layer perceptron classifier outperformed others with an accuracy of 88.55%, establishing its efficacy in lung cancer detection.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In the study detailed by [15], the primary objective was the early diagnosis of lung cancer through the evaluation of classification algorithms. Naive Bayes, support vector machine, decision tree, and logistic regression were applied to lung cancer datasets from UCI and data.world. Logistic regression achieved higher accuracy of 96.9% in the UCI dataset, while support vector machine excelled with a 99.2% accuracy in the data.world dataset.

The research work by [16] aimed to enhance prediction accuracy and Root Mean Square Error (RMSE) of lung cancer patient survival time. A combination of Random Forest classification model and three regression models (general linear regression and gradient-boosted machines) was employed. Random forest exhibited superiority for survival times ≤ 6 and > 24 months, while gradient boosting machine excelled for 7–24 months.

[17] used various well-known classifiers, including support vector machine, C4.5 decision tree, multi-layer perceptron, neural network, and naive Bayes, were applied for the early-stage prediction of lung cancer using a reference dataset from the UCI repository. Ensemble models like random forest and majority voting were also utilized, with the gradient-boosted tree outperforming others, achieving an accuracy of 90%.

The study by [18] aimed to construct a data mining classification model for predicting lung cancer based on the dataset from [39] . Through the CRISP-DM methodology and RapidMiner software, various models and sampling methods were explored, with the artificial neural network algorithm demonstrating superior performance, achieving an accuracy of 92%, recall of 94.2%, and precision of 90.8%.

Finally, in the study by [19], the authors proposed a mechanism for identifying biomarkers for early diagnosis of lung cancer by combining metabolomics mechanisms and machine learning. Utilizing a dataset of 110 lung cancer patients and 43 healthy participants, six specific biomarkers were identified, and Naive Bayes was suggested as the preferred model for early lung tumour prediction. The study demonstrated promising discrimination capabilities with high AUC, sensitivity, and specificity values.

These studies collectively underscore the potential of explainable machine learning in advancing lung cancer research, offering improved model performance and interpretability. They contribute valuable insights into the underlying mechanisms of the disease and the critical factors influencing its development.

III. METHODOLOGY

In this section, we elucidate the dataset upon which our study is grounded and delineate the primary stages of the applied methodology for predicting the risk of lung cancer. Specifically, our approach encompasses the crucial procedure of achieving class balance and ranking features within the balanced dataset. Additionally, we expound upon the frequency of occurrence of nominal features concerning the different lung cancer classes. Furthermore, a comprehensive exposition is provided on the machine learning models employed and the corresponding performance metrics utilized in our analysis.

A. The Description of the Dataset

The dataset named "Lung Cancer," sourced from Kaggle[20], comprises 309 instances with 16 attributes, of which 15 are predictive, and 1 serves as the class attribute denoting lung cancer. The predictive attributes encompass gender, age, smoking habits, presence of yellow fingers, anxiety levels, peer pressure influence, chronic disease status, fatigue levels, allergies, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. This dataset provides a valuable resource for studying the interplay of these diverse factors in the context of lung cancer.

In addition, a synthetic dataset was generated using the Gretel AI platform with specific configuration parameters[21]. The ActGAN model, equipped with privacy filters for handling outliers and ensuring similarity preservation, was employed. The dataset was created based on the code schema_version: "1.0" and includes the same set of attributes as the original Kaggle dataset. The ActGAN model was trained for 10000 epochs, incorporating features such as conditional vector type, reconstruction loss coefficient, and binary encoder configurations. The resulting synthetic dataset offers an augmented resource for exploring the intricacies of lung cancer-related attributes while addressing privacy concerns and preserving data utility.

The configuration parameters specified for the ActGAN model in the synthetic dataset generation process play a crucial role in shaping the characteristics of the generated data. The learning rates for the generator and discriminator, denoted by "generator_lr" and "discriminator_lr" respectively, are set at 0.0001 and 0.00033, with corresponding decay rates specified by "generator_decay" and "discriminator_decay" at 0.000001. These parameters govern the optimization processes, ensuring a balanced training of the generator and discriminator components.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The "batch_size" is automatically determined, allowing the model to dynamically adjust the size of data batches during training. "Discriminator_steps" is set to 1, indicating the number of steps the discriminator takes before updating the generator, contributing to the stability of the adversarial training.

The "binary_encoder_cutoff" is established at 150, determining the threshold for encoding binary variables. In case of missing values in binary-encoded columns, the "binary_encoder_nan_handler" employs the mode to handle these occurrences. The "auto_transform_datetimes" parameter is set to false, indicating that datetime variables are not automatically transformed during the generation process.

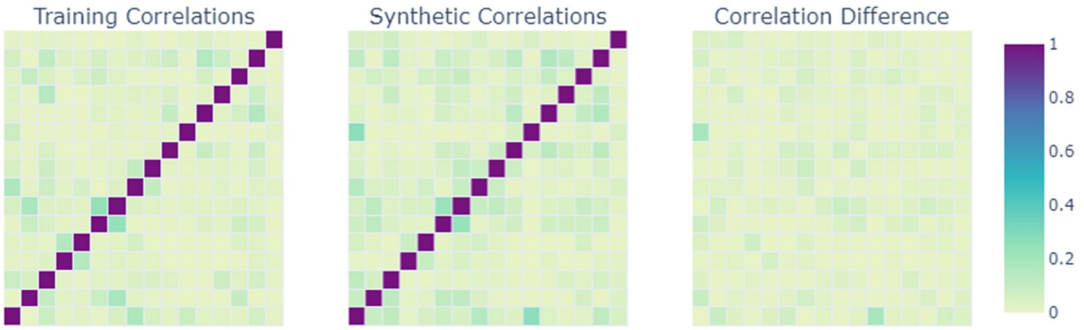


Figure 1: Synthetic dataset correlation with the original dataset from Kaggle.

"Log_frequency" is set to true, facilitating the recording of training progress at regular intervals. The "cbn_sample_size" is configured at 250,000, dictating the size of the conditional batch normalization samples. The model undergoes training for 8,000 epochs ("epochs"), with a "pac" (privacy amplification constant) value of 10 for privacy preservation.

The "data_upsample_limit" is capped at 100, controlling the extent to which data can be upsampled during the generation process. The "conditional_vector_type" is specified as "single_discrete," indicating the type of conditional vectors used in the generation. Parameters related to conditional column selection and mean columns are left as null, allowing for flexibility in conditional generation. The "reconstruction_loss_coef" is set to 1, influencing the weight assigned to the reconstruction loss during training.

Lastly, the "force_conditioning" is set to auto, indicating that the conditioning of the generated data is automatically managed by the model. These parameter configurations collectively tailor the ActGAN model to generate a synthetic dataset that mirrors the statistical characteristics of the original data while incorporating privacy-preserving measures and ensuring robust training dynamics.

Table 1: The Descriptive details of the Attributes in the Dataset.

Attribute	Attribute Description	References
Gender	This attribute discerns the individual's biological sex, categorizing them as either male or female. Understanding gender-specific nuances is crucial in unravelling disparities in lung cancer risk.	[22]
Age (years)	These characteristics record the numerical representation of the participant's age. Age is a pivotal factor, as lung cancer incidence often correlates with advancing age.	[23]
Smoking	This feature denotes whether the participant engages in smoking habits or is a non-smoker. Smoking remains a well-established risk factor for lung cancer, warranting meticulous consideration.	[24]
Yellow fingers	This attribute indicates the presence or absence of yellow fingers in the participant. Yellow fingers can be indicative of prolonged exposure to tobacco, further emphasizing the link with smoking habits.	[25]
Anxiety	This feature reflects the psychological state of the participant, determining if they experience anxiety or not. Exploring mental health aspects contributes to a holistic understanding of factors influencing lung cancer risk.	[26]

Peer pressure	This characteristic captures the participant's response to peer influence, discerning if they feel peer pressure or not. Peer dynamics and societal influences play a role in shaping behaviours, including those related to smoking.	[27]
Chronic disease	This feature expresses the presence or absence of a chronic disease in the participant. Chronic conditions may interact with lung cancer risk, necessitating a comprehensive health assessment.	[28]
Fatigue	This attribute manifests the participant's physical state, indicating whether they suffer from fatigue or not. Fatigue can be both a symptom and a potential contributor to overall health.	[29]
Allergy	This characteristic reveals the presence or absence of allergies in the participant. Exploring the immune system's role adds a layer of complexity to our understanding of lung cancer risk factors.	[30]
Wheezing	This feature declares whether the participant experiences wheezing or not. Wheezing, often associated with respiratory issues, can be indicative of underlying lung health concerns.	[31]
Alcohol	This attribute reflects the participant's alcohol consumption habits, determining if they consume alcohol or not. Alcohol consumption, alongside smoking, contributes to a comprehensive lifestyle assessment.	[32]
Coughing	This feature indicates whether the participant suffers from coughing or not. Chronic coughing can be symptomatic and requires careful consideration in the context of lung health.	[33]
Shortness of breath	This characteristic reveal whether the participant experiences shortness of breath or not. Understanding respiratory symptoms adds crucial information to the risk assessment.	[34]
Swallowing difficulty	This attribute indicates if the participant encounters difficulty swallowing or not. Swallowing difficulties, though multifactorial, may impact the overall well-being of individuals with lung cancer.	[35]
Chest pain	This feature captures whether the participant experiences chest pain or not. Chest pain, while not exclusive to lung cancer, demands exploration as part of a comprehensive health evaluation.	[36]
Lung Cancer	This attribute signifies whether the participant has received a medical diagnosis of lung cancer or not. This critical information serves as the outcome variable for our risk prediction model.	

The dataset under analysis comprises demographic and health-related attributes of individuals, providing a comprehensive snapshot of key factors associated with lung cancer. The gender distribution reveals a predominant representation of males (3,073) compared to females (1,927). Smoking habits present a notable imbalance, with 2,931 individuals identified as smokers (coded as 2) and 2,069 as non-smokers (coded as 1). Concerning psychological aspects, anxiety is prevalent among 3,051 individuals, while 1,949 report no anxiety. Peer pressure showcases a similar distribution, with 2,797 individuals influenced and 2,203 not influenced. Chronic diseases are reported by 2,743 individuals, contrasting with 2,257 without chronic conditions.

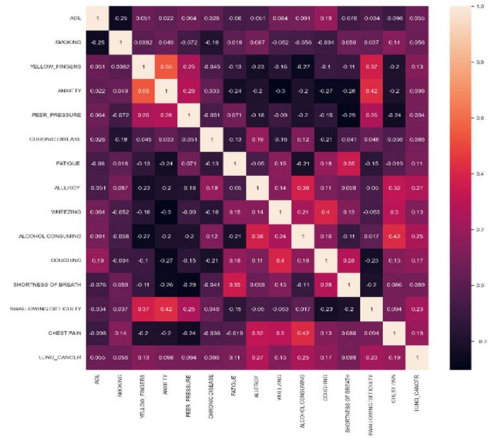


Figure 2: Visualization of the correlation among the attributes in the synthetic dataset.

The dataset reflects varying levels of fatigue and allergies, where 3,201 individuals experience fatigue and 2,133 report allergies. Additionally, symptoms such as wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain exhibit diverse prevalence among the sampled population. Notably, lung cancer instances are prominently represented, with 4,595 individuals diagnosed positively compared to 405 cases without lung cancer.

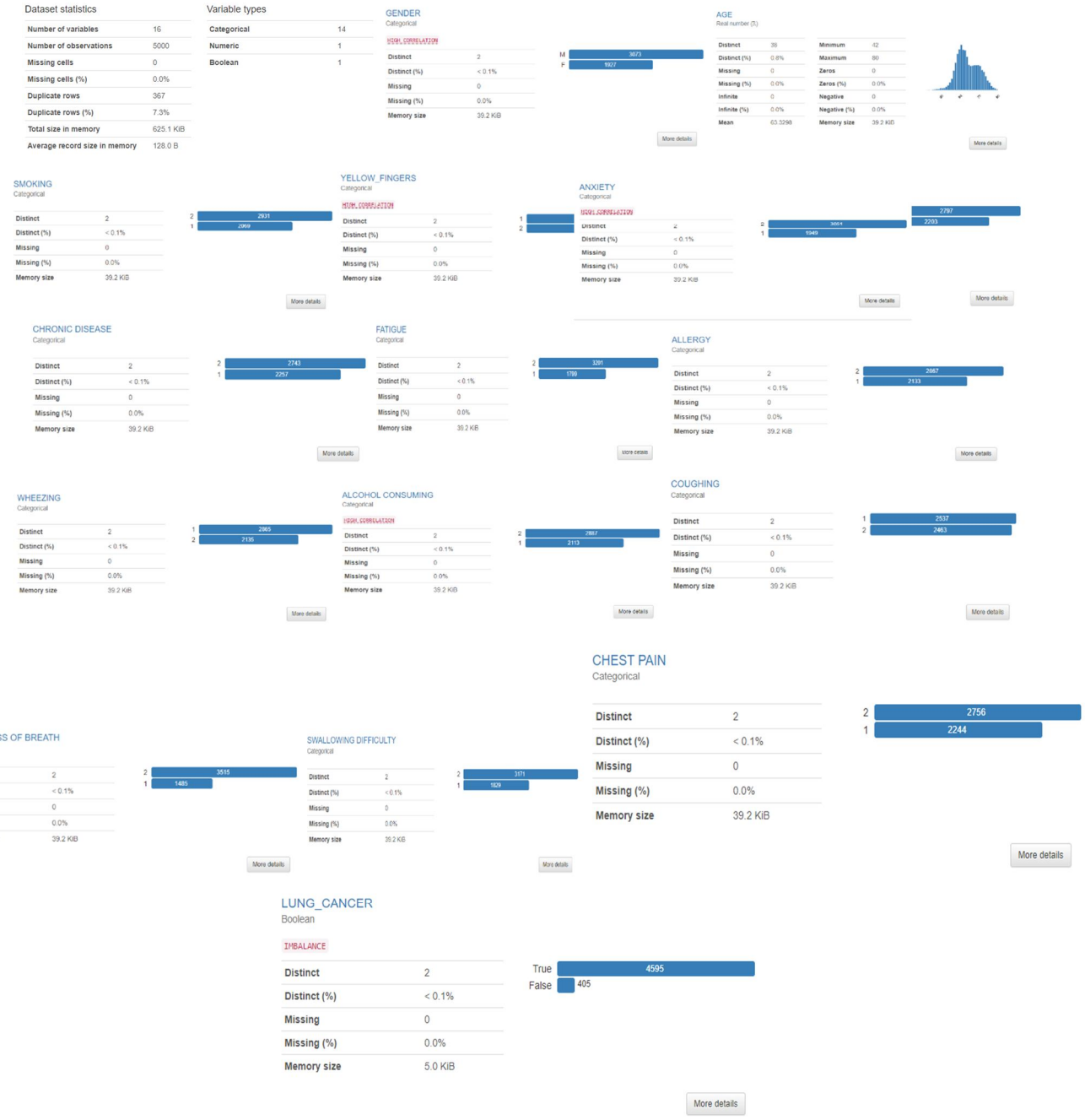


Figure 3: Profile of the Dataset

This dataset, encompassing diverse attributes, offers a nuanced perspective on the interplay of demographic and health factors in the context of lung cancer. The prevalence of certain characteristics provides a foundation for further exploratory analyses and the development of predictive models, contributing to the ongoing efforts in lung cancer research and personalized healthcare interventions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Data Preprocessing

In the preliminary phases of data preprocessing, a pivotal transformation was instituted targeting specific attributes, particularly 'Gender' and 'Lung Cancer,' both characterized by categorical values. This strategic methodology was employed to augment the dataset's adaptability to machine learning algorithms, thereby facilitating proficient pattern recognition and analysis. The ADASYN algorithm, as proposed by [37], serves as a potent solution for addressing class imbalances within datasets. In the realm of lung cancer prediction, the dataset inherently suffers from a disproportionate distribution of instances between the positive (presence of lung cancer) and negative (absence of lung cancer) classes. ADASYN, an adaptive synthetic sampling approach, functions by introducing synthetic instances in regions of the feature space where the minority class (lung cancer instances) is underrepresented. This adaptability is crucial, preventing overfitting by strategically generating synthetic samples in areas characterized by sparse class distribution. This process results in a rebalanced dataset, where synthetic instances have been strategically introduced to rectify the class imbalance, thereby enhancing the learning capacity of the subsequent predictive model.

Another pivotal aspect of the data preprocessing pipeline involves the use of the LabelEncoder from the scikit-learn preprocessing module. This transformation is particularly pertinent for encoding categorical variables into numerical representations, a prerequisite for many machine learning algorithms. This process involves assigning a unique numerical label to each distinct category within a categorical variable. In the 'GENDER' attribute, 'M' be encoded as 1, and 'F' as 2. This transformation ensures that categorical data is expressed in a format compatible with machine learning algorithms, enabling the algorithm to discern patterns and relationships within the data.

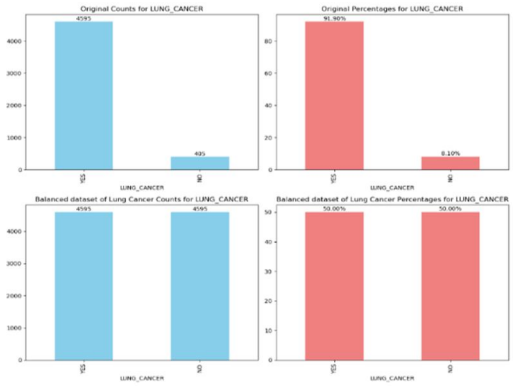


Figure 4: Showing the original dataset (imbalanced dataset) and balanced dataset (count and percentage).

In concert, the amalgamation of ADASYN oversampling and LabelEncoder transformations contributes to a refined and balanced dataset, setting the stage for the development of a robust predictive model for lung cancer risk assessment.

IV. MACHINE LEARNING

In this study, a comprehensive analysis of predictive modelling was conducted utilizing a repertoire of ten distinct machine learning models. The chosen models span various algorithmic domains, encompassing both linear and non-linear methodologies, to comprehensively evaluate prediction performance. The selected models include classical linear models such as Logistic Regression, which is widely employed in binary classification tasks due to its interpretability and efficiency [38]. Decision Tree, another model utilized in this study, operates by recursively partitioning the feature space to create hierarchical decision rules [39]. K-Nearest Neighbors (KNN) relies on proximity-based classification, assigning labels based on the majority class among the nearest neighbours[40]. Bayesian methods were also explored through Gaussian Naïve Bayes and Multinomial Naïve Bayes. Gaussian Naïve Bayes assumes normality in the distribution of features, while Multinomial Naïve Bayes is tailored for discrete data, particularly prevalent in natural language processing tasks[41]. Support Vector Classifier, a powerful tool for both linear and non-linear classification, was employed in this study [42]. Ensemble methods, known for their robustness, were considered, including Random Forest, XGBoost, and Gradient Boost. Random Forest builds a multitude of decision trees and amalgamates their outputs to enhance predictive accuracy [39]. XGBoost, an optimized gradient boosting algorithm, iteratively improves the model's performance by addressing weaknesses in preceding models [43].Gradient Boost, a boosting technique that sequentially combines weak learners, was also integrated into the comparative analysis [44]. Neural networks, specifically Multi-layer Perceptron, were incorporated to explore the complexities captured by deep learning architectures [45].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

This diverse ensemble of machine learning models was chosen to ensure a holistic evaluation of predictive capabilities, considering both linear and non-linear relationships within the data. Each model's distinctive characteristics contribute to a nuanced understanding of their applicability and efficacy in the context of the prediction task at hand.

A. Logistic Regression

Logistic Regression is a statistical method extensively employed for binary classification, where the outcome variable is categorical with two levels, often denoting classes 0 and 1. This model is widely applicable across various domains, including medicine, finance, and social sciences. At its core, Logistic Regression models the probability of an instance belonging to the positive class using the logistic function, ensuring the output is bounded between 0 and 1. The logistic function is defined by the natural logarithm of the odds, incorporating coefficients that represent the relationship between the independent variables and the log-odds of the positive class[46].

B. Decision Tree

The Decision Tree algorithm is a fundamental machine learning technique used for both classification and regression tasks. It operates by recursively partitioning the dataset based on different features, creating a tree-like structure. At each internal node of the tree, a decision is made based on a specific feature, and the dataset is split into subsets. This process continues until a stopping criterion is met, typically when a certain depth is reached, or further splitting does not significantly improve the model's performance. The decision-making process at each node is determined by evaluating the impurity or purity of the data. Common impurity measures include Gini impurity and entropy. The algorithm selects the feature and split point that minimizes impurity, aiming to create homogeneous subsets[47].

C. K-Nearest Neighbor (KNN)

K-Nearest Neighbors (KNN) stands out as a versatile and easy-to-understand machine learning algorithm with applications in both classification and regression tasks. This non-parametric, lazy learning algorithm operates on the principle of proximity, making predictions based on the majority class or average value of 'k' nearest neighbors in the feature space. The simplicity and flexibility of KNN contribute to its widespread use across various domains. It is particularly effective in scenarios where the decision boundary is complex and when dealing with local patterns in the data.[48]

D. Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features. This classification method is particularly powerful for handling continuous data and is widely used in various applications, including natural language processing and medical diagnosis. The underlying principle of GNB is rooted in Bayes' theorem, which calculates the probability of a hypothesis given the observed evidence. In the case of classification, it predicts the probability of a data point belonging to a particular class based on its features. The 'naive' assumption of independence simplifies the computations, as it assumes that the presence of a particular feature in a class is unrelated to the presence of other features[49].

E. Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic classification algorithm that is particularly well-suited for text classification tasks, where the data is represented as word frequency vectors. This algorithm is an extension of the traditional Naive Bayes classifier and assumes that the features follow a multinomial distribution. In MNB, the dataset is typically represented as a term-document matrix, where each row corresponds to a document, each column corresponds to a unique term, and the entries represent the frequency of each term in the document. The algorithm assumes that the term frequencies are generated from a multinomial distribution for each class. One key strength of MNB is its efficiency in handling high-dimensional and sparse datasets, which are common in text classification[50].

F. Support Vector Classifier (SVC)

Support Vector Classifier, or SVC, is a powerful supervised machine learning algorithm used for classification and regression tasks. It belongs to the family of Support Vector Machines (SVM), which are widely employed in various domains due to their robust performance. The primary objective of SVC is to find a hyperplane that best separates the data into different classes.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In a binary classification scenario, this hyperplane aims to maximize the margin between the two classes, which is the distance between the hyperplane and the nearest data points from each class. The algorithm identifies support vectors, which are the data points that influence the position and orientation of the hyperplane[51].

G. Random Forest

Random Forest is a versatile and powerful ensemble learning algorithm widely used for both classification and regression tasks. Ensembles combine multiple base models to improve predictive performance and generalization. Random Forest constructs a multitude of decision trees during training and outputs the mode (classification) or mean (regression) prediction of the individual trees. The strength of Random Forest lies in its ability to mitigate overfitting, enhance accuracy, and handle high-dimensional datasets. The algorithm introduces randomness during the tree-building process by using a subset of features at each split[10].

H. XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized and scalable implementation of the gradient boosting framework, designed for efficiency, speed, and predictive performance. Developed by Tianqi Chen, XGBoost has become a cornerstone algorithm in machine learning competitions and various real-world applications. The key strength of XGBoost lies in its ability to handle both regression and classification tasks, delivering state-of-the-art results with unparalleled speed[52].

I. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a type of artificial neural network designed for supervised learning. It is a versatile and powerful model capable of capturing complex patterns in data, making it suitable for a wide range of applications, including image recognition, natural language processing, and predictive analytics. MLP consists of multiple layers of interconnected nodes, organized into an input layer, one or more hidden layers, and an output layer. Each node, or neuron, in a layer is connected to every node in the subsequent layer, forming a network of interconnected neurons. These connections are associated with weights that are adjusted during the training process to optimize the model's performance. The key strength of MLP lies in its capacity to learn non-linear relationships within data. This is achieved by activation functions, which introduce non-linearities into the model. Popular activation functions include the rectified linear unit (ReLU) and the hyperbolic tangent (tanh)[53].

J. Gradient Boosting

Gradient Boosting is a powerful ensemble machine learning technique that excels in predictive modelling by combining the strengths of multiple weak learners, typically decision trees, to form a robust and accurate predictive model. The fundamental principle behind Gradient Boosting lies in sequentially training new models to correct errors made by the previous ones in the ensemble. This iterative process involves minimizing a predefined loss function, measuring the disparity between actual and predicted values. The algorithm employs gradient descent optimization to iteratively update model parameters, making the model increasingly adept at capturing complex patterns and relationships in the data. A crucial hyperparameter, the learning rate, governs the influence of each weak learner on the overall ensemble, with lower rates enhancing model robustness at the cost of requiring more iterations for convergence[52].

V. EVALUATION METRICS FOR PREDICTIVE MODELLING IN LUNG CANCER DETECTION

To comprehensively assess the performance of the machine learning models employed in this study, a comprehensive set of evaluation metrics, including accuracy, precision, recall, F-Measure, and the Area Under the Curve (AUC), to assess the predictive capabilities of various machine learning models. The evaluation process is facilitated using a confusion matrix, a fundamental tool in classification tasks. This matrix comprises four elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These elements play a pivotal role in calculating the selected metrics and providing insights into the model's behaviour. Accuracy, a foundational metric, offers a holistic view of a model's performance by quantifying the proportion of correctly predicted instances across the entire dataset. While accuracy provides an overall measure, precision, recall, and the F-Measure offer a more nuanced understanding of the model's performance, especially in imbalanced datasets. Precision, a measure of quality, focuses on the accuracy of positive predictions made by the model. A high precision indicates that when the model predicts a positive instance, it is likely to be correct. Conversely, recall, also known as sensitivity or true positive rate, gauges the model's ability to identify all relevant positive instances.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

High recall signifies that the model can effectively capture most of the actual positive cases. The F-Measure, the harmonic mean of precision and recall, provides a balanced assessment, considering both false positives and false negatives [54].

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP}, \quad Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN}, \quad F-Measure = 2\frac{Precision\ X\ Recall}{Precision+Recall} \tag{2}$$

In the realm of medical diagnoses, the consequences of false positives and false negatives can vary significantly. A false positive may lead to unnecessary interventions or treatments, while a false negative could result in a missed opportunity for early intervention.

Therefore, the choice of metrics depends on the specific goals and constraints of the application. The AUC, another critical metric, assesses the model's ability to discriminate between positive and negative instances. It represents the area under the Receiver Operating Characteristic (ROC) curve, a graphical representation of the model's performance across various thresholds. An AUC value closer to 1 indicates superior discriminative ability, while a value of 0.5 suggests random chance. This comprehensive suite of evaluation metrics enables a detailed examination of the strengths and weaknesses of each machine learning model employed in predicting lung cancer.

It goes beyond a binary classification of correct or incorrect predictions and provides a nuanced understanding of the trade-offs between precision and recall. The evaluation of machine learning models in the context of lung cancer prediction involves a multifaceted analysis using various metrics. Accuracy, precision, recall, F-Measure, and AUC collectively contribute to a thorough understanding of a model's performance, allowing stakeholders to make informed decisions based on the specific goals and constraints of the application. As medical applications demand a careful balance between identifying positive cases and minimizing false positives, the choice of evaluation metrics becomes paramount in ensuring the efficacy of predictive models in real-world scenarios[10].

VI. RESULT AND DISCUSSION

A. *Experimental Setup*

Initially, the dataset undergoes a division into training (75%) and testing (25%) subsets. Subsequently, various machine learning methodologies, encompassing feature scaling, Principal Component Analysis (PCA), Random Oversampling (ROS), and hyperparameter tuning, are employed to identify the optimal model with the highest accuracy. Through the application of these diverse techniques, the model selection process is refined, culminating in the identification of the most effective model.

B. *Environmental Setup for the Analysis*

The evaluation of the machine learning (ML) models was conducted within the software Anaconda in the python language known for its versatile libraries catering to data preprocessing, classification, clustering, prediction, and visualization. The experiments were executed on a computer system featuring an Intel(R) Core (TM) i7-6820HQ CPU @ 2.70GHz 2.70 GHz, 16 GB RAM, running Windows 11 Pro with a 64-bit operating system and an x64 processor architecture.

C. *Accuracies of the Model*

Logistic Regression with an AUC of 0.9777, the model excels in discriminating between instances of lung cancer and non-lung cancer, underscoring its effectiveness. The balanced values of Accuracy (0.9221), Precision (0.9202), Recall (0.9240), and F-Measure (0.9221) collectively indicate its reliability in classification tasks. The Decision Tree with an AUC of 0.9565 suggests effective discrimination between classes.

Notably, high values for Precision (0.9395) and Recall (0.9734) contribute to a well-balanced F-Measure (0.9561), emphasizing the model's accuracy in classification tasks. KNN emerges as a robust performer with a notable AUC of 0.9812. The model achieves an excellent balance between Precision (0.9307) and Recall (0.9825), resulting in a high F-Measure (0.9559). The low Error rate (0.0452) signifies effective and reliable classification. Gaussian Naive Bayes demonstrates competitive performance, achieving a balanced trade-off between Precision (0.8982) and Recall (0.8862).

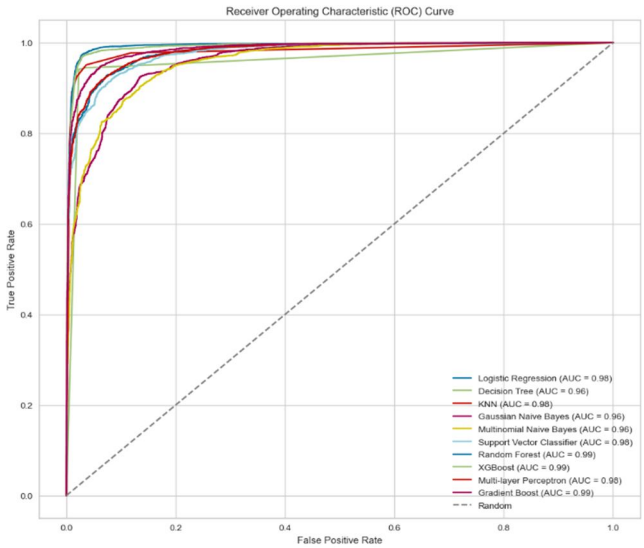


Figure 5: The evaluation of model based on the Area Under the Receiver Operating Characteristic curve.

With an AUC of 0.9584, the model showcases its ability to effectively discriminate between different classes. Random Forest, an ensemble learning method, stands out with the highest AUC of 0.9919. The model achieves a remarkable balance between Precision (0.9605) and Recall (0.9758), resulting in a high overall accuracy. The low Error rate (0.0322) underlines the model's effectiveness in lung cancer prediction. Consistently high AUC values across models affirm their robust discriminatory power, crucial for reliable disease prediction. Each model's unique strengths contribute to the comprehensive evaluation, offering valuable insights into their applicability in the medical domain.

Table 2: Performance analysis of every model. Showing the Area under the curve (AUC), Accuracy, Precision, Recall and F-Measure of the models evaluated.

Models	AUC	Accuracy	Precision	Recall	F-Measure
Logistic Regression	0.9777	0.9221	0.9222	0.9221	0.9221
Decision Tree	0.9573	0.9555	0.9566	0.9563	0.9572
KNN	0.9812	0.9548	0.9563	0.9548	0.9547
Gaussian Naive Bayes	0.9584	0.8931	0.8933	0.8931	0.8931
Multinomial Naive Bayes	0.9571	0.8862	0.8864	0.8862	0.8862
Support Vector Classifier	0.9747	0.9128	0.9136	0.9128	0.9128
Random Forest	0.9924	0.9687	0.9684	0.9687	0.968
XGBoost	0.9912	0.9658	0.9659	0.9658	0.9658
Multi-layer Perceptron	0.9798	0.9219	0.9241	0.9171	0.9167
Gradient Boost	0.9842	0.9394	0.9395	0.9394	0.9394

VII. CONCLUSION

The proposed methodology in this study revolves around utilizing a dataset comprising features that encapsulate human habits such as smoking and alcohol consumption, alongside signs/symptoms considered as potential risk factors for lung cancer patients. It's noteworthy that these signs may not exclusively pertain to lung cancer. Unlike other types of cancers, lung cancer remains imperceptible to the naked eye, often exhibiting symptoms concomitant with other diseases. Common manifestations include allergies, asthma, shortness of breath, and coughing. In our approach, we opted to train various classifiers on diverse risk factors associated with these symptoms to accurately discern the class label (Lung Cancer or Non-Lung Cancer) of an unknown instance, thereby estimating the associated risk. Even in the absence of overt disease manifestation, adopting practices like risk-factor monitoring and follow-up clinical examination proves beneficial for lung cancer management, potentially mitigating adverse effects through early diagnosis.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Clinical examination and identification of lung cancer typically involve imaging techniques such as X-ray, CT, PET-CT, and MRI scans of the patient's chest. Consequently, leveraging the considered dataset alongside features derived from lung images holds considerable promise for the early diagnosis and staging of the disease[55]. It's imperative to note that our study exclusively focuses on discerning the presence or absence of lung cancer, framing it as a binary classification problem. While machine learning offers multi-class classification strategies for cancer stage identification, such as one vs. one (OVO) and one vs. all (OVA) methods, the nature of our dataset precludes the adoption of such approaches[56].

Undoubtedly, machine learning has emerged as a crucial tool for medical practitioners and clinicians in the realms of early screening, prediction, and prognosis of various diseases. Researchers have invested substantial efforts in accessing medical information, collecting data through questionnaires, or generating datasets in laboratories to bolster healthcare analytics. It is pertinent to acknowledge a limitation in our study. This research paper drew from a publicly available dataset and with the help of that created a synthetic dataset based on which the all the analysis and model was trained for prediction, lacking the depth and diversity that might have been afforded by data sourced directly from a hospital unit or institute. Furthermore, obtaining access to sensitive medical data remains challenging due to privacy concerns. Nonetheless, the dataset utilized in our study boasted beneficial features that facilitated the derivation of reliable and accurate research results.

A. Funding
This research received no external funding.

B. Data Availability Statement
https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.
https://gretel.ai/

C. Conflicts of Interest
The authors declare no conflict of interest.

REFERENCES

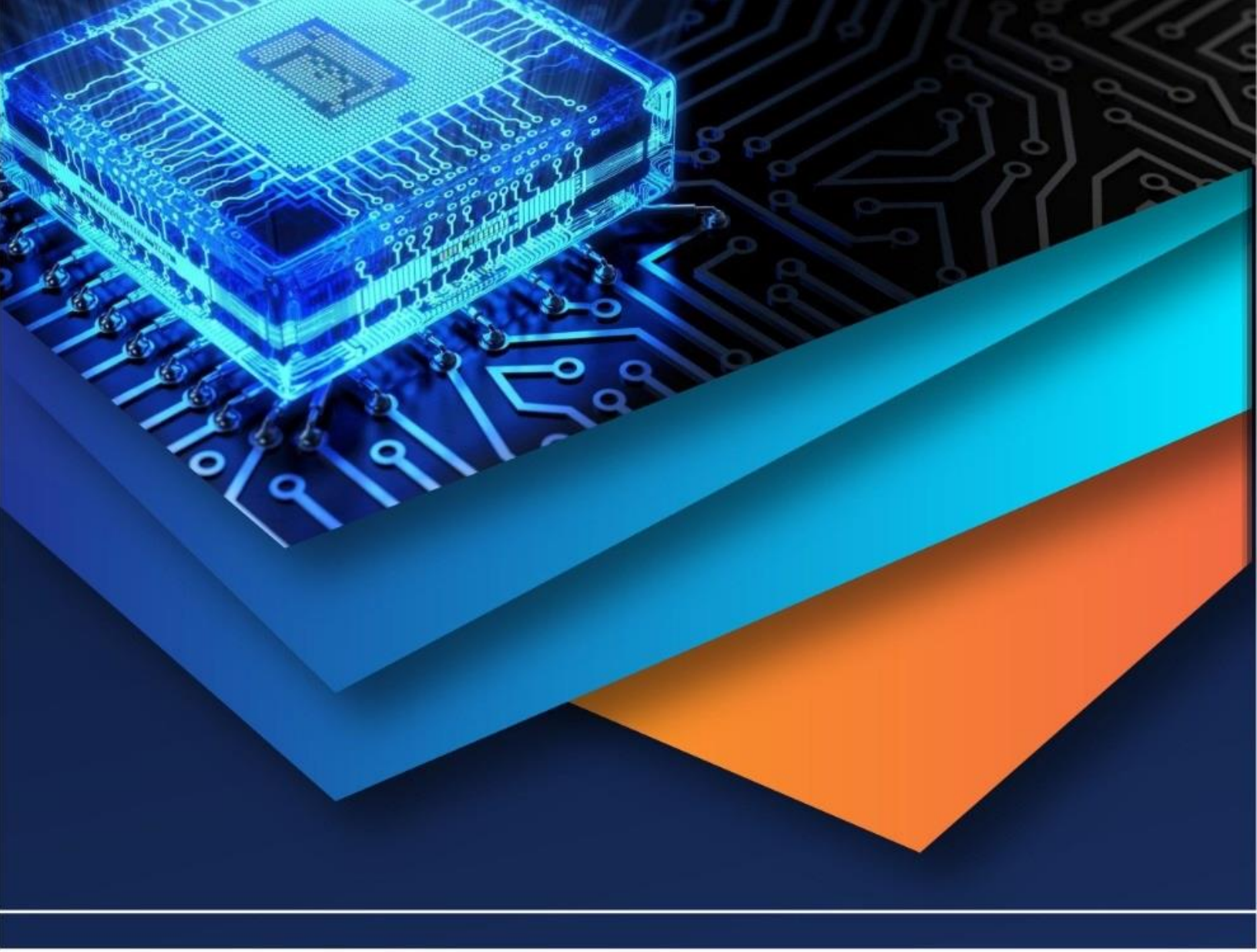
[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA Cancer J Clin, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
[2] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nat Med, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
[3] A. Esteva et al., “A guide to deep learning in healthcare,” Nat Med, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.
[4] A. Yaqoob, R. Musheer Aziz, and N. K. verma, “Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review,” Hum-Cent Intell Syst, vol. 3, no. 4, pp. 588–615, Dec. 2023, doi: 10.1007/s44230-023-00041-3.
[5] P. Goldstraw et al., “The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer,” J Thorac Oncol, vol. 11, no. 1, pp. 39–51, Jan. 2016, doi: 10.1016/j.jtho.2015.09.009.
[6] M. M. Sobhan Ananda Mohan, “Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer,” NA, vol. NA, no. NA, p. NA-NA, 2022, doi: 10.1101/2022.10.13.512119.
[7] B. A. Alsinglawi Osama; Alorjani, Mohammed; Mubin, Omar; Alnajjar, Fady; Novoa, Mauricio; Darwish, Omar, “An explainable machine learning framework for lung cancer hospital length of stay prediction.,” Scientific reports, vol. 12, no. 1, pp. 607-NA, 2022, doi: 10.1038/s41598-021-04608-7.
[8] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, “Explainable Machine Learning for Lung Cancer Screening Models,” Applied Sciences, vol. 12, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/app12041926.
[9] J. Duell, X. Fan, B. Burnett, G. Aarts, and S.-M. Zhou, “A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records,” in 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Jul. 2021, pp. 1–4. doi: 10.1109/BHI50953.2021.9508618.
[10] E. T. Dritsas Maria, “Lung Cancer Risk Prediction with Machine Learning Models,” Big Data and Cognitive Computing, vol. 6, no. 4, pp. 139–139, 2022, doi: 10.3390/bdcc6040139.
[11] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” in 2022 IEEE World AI IoT Congress (AIIoT), Jun. 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.
[12] R. Patra, “Prediction of Lung Cancer Using Machine Learning Classifier,” in Communications in Computer and Information Science, vol. 1235, NA vols., 2020, pp. 132–142. doi: 10.1007/978-981-15-6648-6_11.
[13] J. A. K. Sim Young A; Kim, Ju Han; Lee, Jong Mog; Kim, Moon Soo; Shim, Young Mog; Zo, Jae Ill; Yun, Young Ho, “The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning,” Scientific reports, vol. 10, no. 1, pp. 10693-NA, 2020, doi: 10.1038/s41598-020-67604-3.
[14] G. A. P. G. Singh Pradeep Kumar, “Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans,” Neural Computing and Applications, vol. 31, no. 10, pp. 6863–6877, 2018, doi: 10.1007/s00521-018-3518-x.

[15] R. P.R., R. A. S. Nair, and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Feb. 2019, pp. 1–4. doi: 10.1109/ICECCT.2019.8869001.
[16] J. A. ; F. Bartholomai Hermann B., "ISSPIT - Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques," Proceedings of the ... IEEE International Symposium on Signal Processing and Information Technology. IEEE International Symposium on Signal Processing and Information Technology, vol. 2018, no. NA, pp. 632–637, 2018, doi: 10.1109/isspit.2018.8642753.
[17] M. I. B. Faisal Saba; Khan, Zain Sikandar; Khan, Farhan Hassan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), vol. NA, no. NA, p. NA-NA, 2018, doi: 10.1109/iceest.2018.8643311.
[18] E. Vieira, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data Mining Approach to Classify Cases of Lung Cancer," in Trends and Applications in Information Systems and Technologies, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2021, pp. 511–521. doi: 10.1007/978-3-030-72657-7_49.
[19] Y. Xie et al., "Early lung cancer diagnostic biomarker discovery by machine learning methods," Translational Oncology, vol. 14, no. 1, p. 100907, Jan. 2021, doi: 10.1016/j.tranon.2020.100907.
[20] "Lung Cancer." Accessed: Mar. 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[21] "Gretel Console." Accessed: Mar. 17, 2024. [Online]. Available: https://console.gretel.ai/dashboard/
[22] C. Stapelfeld, C. Dammann, and E. Maser, "Sex-specificity in lung cancer risk," Int J Cancer, vol. 146, no. 9, pp. 2376–2382, May 2020, doi: 10.1002/ijc.32716.
[23] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, "The epidemiology of lung cancer," Transl Lung Cancer Res, vol. 7, no. 3, pp. 220–233, Jun. 2018, doi: 10.21037/tlcr.2018.05.06.
[24] L. M. O'Keeffe, G. Taylor, R. R. Huxley, P. Mitchell, M. Woodward, and S. A. E. Peters, "Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis," BMJ Open, vol. 8, no. 10, p. e021611, Oct. 2018, doi: 10.1136/bmjopen-2018-021611.
[25] B. Al-Bander, Y. A. Fadil, and H. Mahdi, "Multi-Criteria Decision Support System for Lung Cancer Prediction," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1076, no. 1, p. 012036, Feb. 2021, doi: 10.1088/1757-899X/1076/1/012036.
[26] T. Hu, J. Xiao, J. Peng, X. Kuang, and B. He, "Relationship between resilience, social support as well as anxiety/depression of lung cancer patients: A cross-sectional observation study," Journal of Cancer Research and Therapeutics, vol. 14, no. 1, p. 72, 2018, doi: 10.4103/jcrt.JCRT_849_17.
[27] C. T. Leshargie et al., "The impact of peer pressure on cigarette smoking among high school and university students in Ethiopia: A systemic review and meta-analysis," PLoS One, vol. 14, no. 10, p. e0222572, 2019, doi: 10.1371/journal.pone.0222572.
[28] M. B. Schabath and M. L. Cote, "Cancer Progress and Priorities: Lung Cancer," Cancer Epidemiology, Biomarkers & Prevention, vol. 28, no. 10, pp. 1563–1579, Oct. 2019, doi: 10.1158/1055-9965.EPI-19-0221.
[29] A. Avancini et al., "Physical Activity and Exercise in Lung Cancer Care: Will Promises Be Fulfilled?," Oncologist, vol. 25, no. 3, pp. e555–e569, Mar. 2020, doi: 10.1634/theoncologist.2019-0463.
[30] E. D. Kantor, M. Hsu, M. Du, and L. B. Signorello, "Allergies and Asthma in Relation to Cancer Risk," Cancer Epidemiol Biomarkers Prev, vol. 28, no. 8, pp. 1395–1403, Aug. 2019, doi: 10.1158/1055-9965.EPI-18-1330.
[31] N. A. Alsharairi, "The Effects of Dietary Supplements on Asthma and Lung Cancer Risk in Smokers and Non-Smokers: A Review of the Literature," Nutrients, vol. 11, no. 4, p. 725, Mar. 2019, doi: 10.3390/nu11040725.
[32] D. R. Brenner et al., "Alcohol consumption and lung cancer risk: A pooled analysis from the International Lung Cancer Consortium and the SYNERGY study," Cancer Epidemiol, vol. 58, pp. 25–32, Feb. 2019, doi: 10.1016/j.canep.2018.10.006.
[33] A. S. M. Harle et al., "Cough in Patients With Lung Cancer: A Longitudinal Observational Study of Characterization and Clinical Associations," Chest, vol. 155, no. 1, pp. 103–113, Jan. 2019, doi: 10.1016/j.chest.2018.10.003.
[34] M. Phillips, T. L. Bauer, and H. I. Pass, "A volatile biomarker in breath predicts lung cancer and pulmonary nodules," J. Breath Res., vol. 13, no. 3, p. 036013, Jun. 2019, doi: 10.1088/1752-7163/ab21aa.
[35] G. C. Brady, J. W. G. Roe, M. O' Brien, A. Boaz, and C. Shaw, "An investigation of the prevalence of swallowing difficulties and impact on quality of life in patients with advanced lung cancer," Support Care Cancer, vol. 26, no. 2, pp. 515–519, Feb. 2018, doi: 10.1007/s00520-017-3858-6.
[36] K. Malinowska, "The relationship between chest pain and level of perioperative anxiety in patients with lung cancer.," Polski przeglad chirurgiczny, vol. 90, no. 2, pp. 23–27, 2018, doi: 10.5604/01.3001.0011.7490.
[37] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.
[38] D. W. H. Jr, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression. John Wiley & Sons, 2013.
[39] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. New York: Chapman and Hall/CRC, 2017. doi: 10.1201/9781315139470.
[40] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
[41] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," Machine Learning, vol. 29, no. 2, pp. 103–130, Nov. 1997, doi: 10.1023/A:1007413511361.
[42] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[43] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[44] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[45] H. Rodrigo and C. P. Tsokos, "Artificial Neural Network Model for Predicting Lung Cancer Survival," Journal of Data Analysis and Information Processing, vol. 5, no. 1, Art. no. 1, Dec. 2016, doi: 10.4236/jdaip.2017.51003.
[46] H. Yuan et al., "Application of logistic regression and convolutional neural network in prediction and diagnosis of high-risk populations of lung cancer," European Journal of Cancer Prevention, vol. 31, no. 2, p. 145, Mar. 2022, doi: 10.1097/CEJ.0000000000000684.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[47] M. Sherafatian and F. Arjmand, “Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data,” Oncol Lett, vol. 18, no. 2, pp. 2125–2131, Aug. 2019, doi: 10.3892/ol.2019.10462.
[48] N. Maleki, Y. Zeinali, and S. T. A. Niaki, “A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection,” Expert Systems with Applications, vol. 164, p. 113981, Feb. 2021, doi: 10.1016/j.eswa.2020.113981.
[49] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, “Cancer Classification Using Gaussian Naive Bayes Algorithm,” in 2019 International Engineering Conference (IEC), Jun. 2019, pp. 165–170. doi: 10.1109/IEC47844.2019.8950650.
[50] S. Nageswaran et al., “Lung Cancer Classification and Prediction Using Machine Learning and Image Processing,” Biomed Res Int, vol. 2022, p. 1755460, Aug. 2022, doi: 10.1155/2022/1755460.
[51] S. HUANG, N. CAI, P. P. PACHECO, S. NARANDES, Y. WANG, and W. XU, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics,” Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, Dec. 2017, doi: 10.21873/cgp.20063.
[52] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, “XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer,” Journal of Pathology Informatics, vol. 14, p. 100307, Jan. 2023, doi: 10.1016/j.jpi.2023.100307.
[53] T. M et al., “Lung cancer diagnosis based on weighted convolutional neural network using gene data expression,” Sci Rep, vol. 14, no. 1, p. 3656, Feb. 2024, doi: 10.1038/s41598-024-54124-7.
[54] M. L. Zaman Chung-Horng, “NOMS - Evaluation of machine learning techniques for network intrusion detection,” NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, vol. NA, no. NA, pp. 1–5, 2018, doi: 10.1109/noms.2018.8406212.
[55] A. Vial et al., “The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review,” Translational Cancer Research, vol. 7, no. 3, Jun. 2018, doi: 10.21037/tcr.2018.05.02.
[56] B. Zhang et al., “Machine learning in chronic obstructive pulmonary disease,” Chin Med J (Engl), vol. 136, no. 5, pp. 536–538, Mar. 2023, doi: 10.1097/CM9.0000000000002247.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)