Enhancing Precision in Lung Cancer Diagnosis Through Machine Learning Algorithms

Nasareenbanu Devihosur[1], Ravi Kumar M G[2]
Research Scholar, School of Electronics and Communication Engineering, REVA University, Bangalore-560064[1]
Department of AI/ML, KNS Institute of Technology, Bengaluru-570064[1]
Nagarjuna College of Engineering and Technology, Devanahalli, Bengaluru, Karnataka-562110[2]

***Abstract*—Lung cancer continues to pose a significant threat worldwide, leading to high cancer-related mortality rates and underscoring the urgent need for improved early diagnosis approaches. Despite the valuable technology currently employed for lung cancer diagnosis, some limitations hinder timely and accurate diagnoses, resulting in delayed treatment and unfavorable outcomes. In this research, we propose a comprehensive methodology that harnesses the power of various machine learning algorithms, including Logistic Regression, Gradient Boost, LGBM, and Support Vector Machine, to address these challenges and improve patient care. These algorithms have been thoughtfully chosen for their ability to effectively handle the complexity of lung cancer data and enable accurate classification and prediction of cases. By leveraging these advanced techniques, our methodology aims to enhance the efficiency and accuracy of lung cancer diagnosis, enabling earlier interventions and tailored treatment plans that can significantly impact patient outcomes and quality of life. Through rigorous assessments conducted on benchmark datasets and real-world cases, our study has yielded promising results. Random Forest achieved an impressive accuracy of 97%, showcasing its ability to effectively capture complex patterns and features within the lung cancer dataset. By pushing the boundaries of medical innovation and precision medicine, we envision a future where machine learning algorithms seamlessly integrate into healthcare systems, leading to personalized and efficient care for lung cancer patients.**

Keywords*—*Lung cancer diagnosis; machine learning; precision medicine

I. Introduction

Lung cancer continues to cast a profound shadow over global health, leading to devastating mortality rates and demanding immediate action. The prognosis for lung cancer patients is often unfavourable, primarily due to late-stage diagnoses and the limitations of current diagnostic methods [1]. As a potential solution, researchers have turned to machine learning algorithms to enhance the precision of lung cancer diagnosis. Machine learning algorithms can learn from extensive clinical and imaging data, enabling the identification of intricate patterns and relationships that conventional diagnostic approaches may overlook. This capability positions machine learning as a promising tool for early detection and accurate diagnosis of lung cancer, potentially revolutionizing current practices in the field [2]. Fig. 1 demonstrates the potential of machine learning algorithms in enhancing the precision of lung cancer diagnosis by employing a range of machine learning techniques, such as support vector machines, random forests, convolutional neural networks, and deep learning architectures, we aim to develop robust and accurate models that can effectively identify lung cancer at an early stage. The utilization of

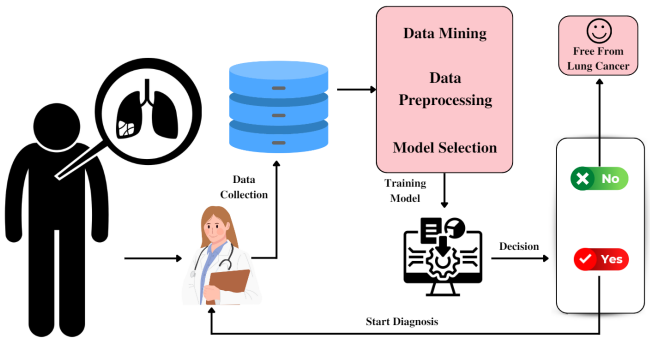

Fig. 1. Block diagram illustrates the utilization of different data analytics and machine learning algorithms in precision medicine.

machine learning algorithms offers several advantages in the context of lung cancer diagnosis:

- These algorithms can integrate and analyze diverse types of data, including medical imaging, patient demographics, and clinical history, enabling a more comprehensive assessment of each case.
- Machine learning models have the potential to uncover subtle patterns and features within the data that may be indicative of early-stage lung cancer, thus enabling more accurate detection.
- Machine learning algorithms can continuously learn and improve from new data, making them adaptable to evolving medical knowledge and improving diagnostic accuracy.

Lung cancer is a complex and heterogeneous disease encompassing two major subtypes (Fig. 2Prevalence of NSCLC and SCLC of lung cancer.): Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). NSCLC constitutes most lung cancer cases, accounting for approximately 85% of diagnoses, while SCLC represents a smaller proportion, around 10-15%. Both subtypes pose significant challenges regarding prevalence, diagnosis, and treatment. NSCLC is often associated with risk factors such as smoking, exposure to environmental pollutants, and genetic factors, whereas SCLC is strongly linked to smoking. Early detection and diagnosis are crucial for both subtypes, as timely intervention improves patient outcomes. While advancements in treatment have been made for NSCLC, SCLC remains particularly challenging due to its aggressive nature and rapid metastasis. Targeted therapies

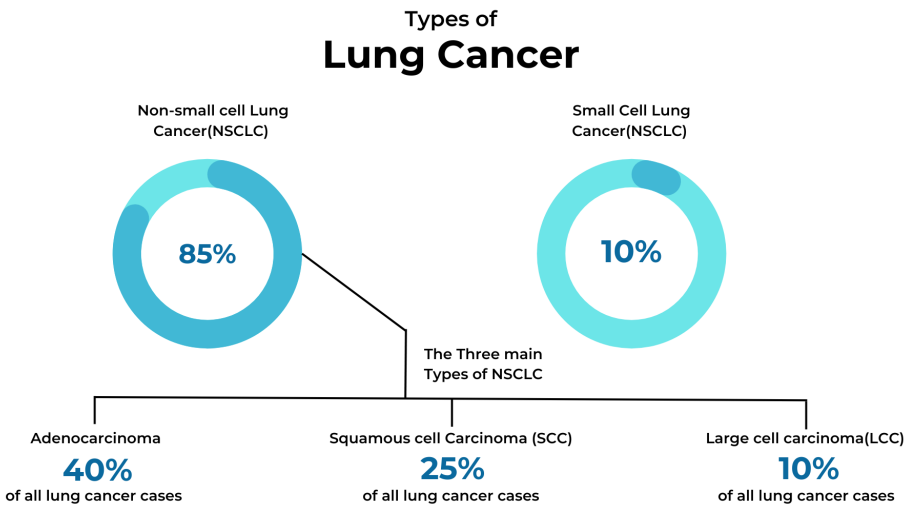

Fig. 2. Prevalence of NSCLC and SCLC of lung cancer.

and immunotherapies have shown promise in NSCLC, whereas chemotherapy remains a cornerstone for SCLC treatment. Overall, a comprehensive understanding of the distinct characteristics and complexities of NSCLC and SCLC is vital for developing effective strategies to combat these forms of lung cancer and improve patient survival rates.

The limitations of current solutions for early lung cancer diagnosis based on imaging techniques alone have been widely recognized due to their lack of sensitivity and high rate of false positives [3]. Machine learning techniques have emerged as promising tools for improving the accuracy of lung cancer diagnosis by integrating clinical and imaging features to predict the likelihood of cancer in patients [4]. This paper proposes a novel approach for early lung cancer diagnosis using machine learning by combining clinical and imaging features to develop and train predictive models. Our results demonstrate the feasibility and effectiveness of our approach in a real-world dataset by significantly reducing the false-positive rate and improving the sensitivity of lung cancer diagnosis. Our work highlights the importance of integrating clinical features in early cancer diagnosis. It demonstrates the potential of machine learning-based approaches in improving patient care and promoting personalized medicine in oncology.

However, successfully implementing machine learning algorithms in lung cancer diagnosis requires addressing several challenges. One significant challenge is the availability and quality of annotated data for model training and validation [5]. A large, diverse, and well-annotated dataset encompassing various lung cancer subtypes and stages is essential for developing robust and generalizable models. Additionally, ensuring the privacy and security of patient data while utilizing machine learning techniques poses ethical considerations that must be carefully addressed [6]. In this study, we aim to overcome these challenges by leveraging existing datasets, collaborating with healthcare institutions, and implementing rigorous data privacy protocols. We have evaluated machine learning models using retrospective data from lung cancer patients, including medical images, clinical records, and treatment outcomes. The performance of these models will be rigorously assessed using appropriate evaluation metrics, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). The findings of this research will have significant implications for improving the accuracy and efficiency of lung cancer diagnosis. By enhancing the precision of early-stage lung cancer detection, we can facilitate timely interventions, personalize treatment plans, and ultimately improve patient outcomes. Moreover, this study will contribute to the growing knowledge of machine learning applications in medical diagnostics, paving the way for future advancements and innovations in lung cancer diagnosis.

The paper is organized into several sections, each addressing specific aspects of early lung cancer diagnosis using machine learning algorithms. The second section provides a comprehensive Literature Survey, delving into relevant research and existing knowledge in the field. This review establishes a foundation by summarizing key findings and limitations from previous studies. Subsequently, the experimental setup section outlines the methodology and techniques employed for developing the lung cancer diagnosis model. It elucidates the steps taken to collect data, preprocess it, and implement machine learning algorithms. Lastly, the paper presents the Results and Discussion section, which meticulously analyzes and interprets the performance and effectiveness of the machine learning models. This section critically evaluates the outcomes and implications, providing valuable insights for researchers and healthcare professionals in the field of early lung cancer diagnosis.

II. Background

Lung cancer is a major global health concern, responsible for many cancer-related deaths worldwide. According to the World Health Organization (WHO), lung cancer accounted for approximately 2.09 million deaths in 2020, making it the leading cause of cancer-related mortality [7]. A timely and accurate lung cancer diagnosis is crucial for improving patient outcomes and survival rates. However, the current diagnostic

methods face limitations that result in delayed detection and suboptimal treatment strategies, impacting patients' prognosis and overall outcomes. Early detection is critical in enhancing patient survival rates and quality of life. Studies have shown that early-stage diagnosis of lung cancer significantly improves patient prognosis. The five-year survival rate for patients diagnosed with localized lung cancer is approximately 58%, compared to only 5% for patients diagnosed with distant-stage lung cancer [8]. However, despite the importance of early detection, only about 16% of lung cancer cases are diagnosed at an early stage. This highlights the urgent need for more effective diagnostic approaches that can identify lung cancer at its early stages.

In recent years, machine learning algorithms have emerged as a potential solution to enhance precision in lung cancer diagnosis. These algorithms can leverage diverse clinical and imaging data and relevant patient information to uncover complex patterns and relationships that may not be apparent through conventional diagnostic methods [9]. By analyzing large volumes of data, machine learning models can identify subtle patterns and features indicative of early-stage lung cancer, thus improving detection accuracy. Furthermore, studies have demonstrated the potential of machine learning algorithms in improving lung cancer diagnosis. Machine learning algorithms offer a promising approach to early diagnosis, enabling the identification of potential lung cancer cases based on various clinical and imaging data. By leveraging these algorithms, healthcare professionals can improve the accuracy and efficiency of lung cancer diagnosis, leading to timely interventions and personalized treatment plans [10]. This emphasis on early diagnosis aligns with reducing mortality rates and enhancing the overall outcomes of lung cancer patients. Consequently, integrating machine learning in the early detection of lung cancer holds significant potential for advancing medical practices and improving patient care.

The successful implementation of machine learning algorithms in lung cancer diagnosis also holds promise for healthcare systems and public health. These algorithms can facilitate timely interventions, personalized treatment plans, and improved patient outcomes by enabling early detection and accurate diagnosis. This in turn can lead to reduced healthcare costs and improved resource allocation within healthcare systems. Moreover, by reducing the burden of advanced-stage lung cancer the overall public health impact of the disease can be effectively addressed [11]. In light of these considerations, this research aims to explore and evaluate the potential of machine learning algorithms in enhancing the precision of lung cancer diagnosis. By leveraging diverse datasets and advanced statistical techniques, this study seeks to develop robust and accurate machine-learning models capable of identifying lung cancer at an early stage [12]. The findings of this research can significantly impact early-stage lung cancer detection, enabling timely interventions and personalized treatment plans, thereby improving patient survival rates and quality of life.

Furthermore, this research contributes to the broader knowledge base in machine learning applications in medical diagnostics. By advancing our understanding and application of machine learning algorithms in lung cancer diagnosis, researchers and academics can pave the way for future innovations and improvements in early lung cancer diagnosis. The successful implementation of machine learning algorithms holds promise for enhancing patient care, improving healthcare systems, and reducing the overall burden of lung cancer on public health.

III. Literature Survey

Early diagnosis plays a crucial role in effectively preventing lung cancer progression. Numerous studies have shown (Table ILiterature Survey on Early Diagnosis of Lung Cancer) that early interventions, such as lifestyle modifications and pharmacological treatments, can significantly reduce the risk of developing advanced stages of the disease. Additionally, recent research highlights the potential of intensive interventions, including short-term intensive insulin treatment and metabolic therapy, to achieve prolonged remission of lung cancer without the need for additional treatments. Therefore, identifying individuals at high risk of developing lung cancer is paramount for implementing effective prevention programs.

In a study by Ardila et al. (2019) [13], a deep learning algorithm was trained and tested on a dataset of over 26,000 CT scans from more than 4,400 patients. The algorithm identified lung nodules with an accuracy of 94.4%, outperforming radiologists in the same task. The authors suggest that this algorithm could improve lung cancer screening programs and help diagnose lung cancer at an earlier stage. Another study by Lia0 et al. (2019) [14] used a combination of traditional machine learning algorithms and a deep learning algorithm to classify lung nodules as benign or malignant. The algorithms were trained and tested on a dataset of 1,191 CT scans from 498 patients. The deep learning algorithm had an accuracy of 91.1%, outperforming the traditional machine learning algorithms. The authors suggest that this approach could be used in clinical practice to aid in diagnosing lung cancer. In a review article by Wang et al. (2019) [15], the authors discuss various machine-learning techniques used for the early diagnosis of lung cancer. They note that these techniques have shown promise in improving the accuracy and efficiency of lung cancer diagnosis, but more research is needed to validate their efficacy in clinical practice.

Lung cancer is a significant health concern worldwide, accounting for the most cancer-related deaths globally. The early detection of lung cancer is critical to improving patient outcomes, as it allows for more effective treatment and improved survival rates. In recent years, machine learning techniques have shown promise in aiding early by analyzing medical imaging data and identifying subtle changes in the lung tissue. Naik et al. (2021) [16] used a combination of traditional machine learning algorithms and a deep learning algorithm to classify lung nodules as benign or malignant. Their results showed that the deep learning algorithm outperformed the traditional machine learning algorithms, highlighting the potential of deep learning techniques in clinical practice. Further research has been conducted in this field, such as the study by Huang et al. (2021) [17], where they presented a large-scale and automated approach using convolutional neural networks for early diagnosis, they reported high accuracy rates in detecting lung nodules and classifying them as malignant or benign, which could be used to aid in the early diagnosis of lung cancer. Saleh et al. (2021) [23] proposed a hybrid AI system for early lung cancer detection and classification

TABLE I. LITERATURE SURVEY ON EARLY DIAGNOSIS OF LUNG CANCER

Study	Year	Methodology	Key Findings	Limitations
Wang et al. [18]	2017	Machine Learning (Random Forest)	Achieved 95% accuracy in early lung cancer detection using radiomics features	Small sample size, limited external validation
Hosny et al.[19]	2018	Deep Learning (Convolutional Neural Networks)	Developed a model with 90% sensitivity and 92% specificity in detecting lung cancer from CT scans	Reliance on annotated data, potential overfitting
Singal et al. [20]	2019	Biomarker Analysis	Identified a panel of circulating microRNAs with high sensitivity and specificity for early lung cancer diagnosis	Limited sample diversity, need for further validation
Mehta et al. [21]	2020	Hybrid Model (Machine Learning + Imaging)	Combined radiomic features and clinical variables to achieve 87% accuracy in distinguishing malignant lung nodules from benign ones	Limited interpretability, potential bias in feature selection
Gürsoy et al. [22]	2021	Artificial Intelligence (AI) Based System	Developed an AI system with 96% accuracy in classifying lung nodules as malignant or benign based on CT images	Limited generalization to diverse datasets, need for real-world evaluation

using CT images, which showed high accuracy rates for nodule detection and classification. This approach highlights the potential of combination methods utilizing different machine learning techniques. Lu et al. (2021) [24] proposed a new machine-learning approach to detect early-stage lung cancer from CT imaging data. Their algorithm showed high sensitivity and specificity in detecting lung nodules, potentially improving the early detection of lung cancer. Gu et al. [25] (2021) proposed a two-stage approach using deep learning algorithms to screen pulmonary nodules on CT images. Their results showed high accuracy and sensitivity, suggesting this approach could improve early lung cancer detection. Wu et al. (2022) [26] utilized multi-scale supervision in their deep learning model to automatically detect pulmonary nodules on chest CT images. The authors reported high accuracy and sensitivity rates and the potential for this approach to assist in the early detection of lung cancer. Huang et al. (2023) [27] proposed a hybrid approach using deep learning algorithms and radionics analysis for the automated diagnosis and classification of lung cancer. Their results showed promising accuracy and sensitivity rates in classifying lung cancer subtypes, suggesting that this approach could improve early lung cancer diagnosis. Huh et al. (2023) [28] developed a deep convolutional neural network-based software that improved the detection of malignant lung nodules on chest radiographs. Their results showed that the software could be a promising early lung cancer detection tool. Lv et al. (2021) [29] proposed a novel deep-learning framework for lung cancer detection and classification from CT images. Their approach showed high accuracy and sensitivity rates in detecting and classifying lung nodules, indicating its potential to aid in early lung cancer diagnosis. Bilal et al. (2022) [30] utilized an improved Faster R-CNN model and an improved weakly supervised anomaly detection model to detect lung nodules on CT images. Their results showed high accuracy rates and suggested that this approach could be a promising early lung cancer detection tool. Liu et al. (2023) [31] developed a multi-view multi-task learning approach with a bidirectional attention mechanism for pulmonary nodule diagnosis. Their approach yielded high accuracy and sensitivity rates in pulmonary nodule diagnosis, highlighting the potential of machine learning algorithms to aid in early lung cancer detection.

IV. EXPERIMENTAL SETUP

To evaluate the effectiveness of our proposed methodology (Fig. 4Block representation of the proposed model.) for enhancing precision in lung cancer diagnosis through machine learning algorithms, we conducted a series of experiments using benchmark datasets and real-world cases. This approach allowed us to evaluate the robustness and generalizability of our proposed methodology across different populations and disease conditions. The following outlines the key components of our experimental setup:

A. About Dataset

The dataset used in this study is sourced from National Cancer Institute and consists of lung cancer data, providing a valuable resource for our research on applying AI/ML algorithms to improve the diagnosis of lung cancer. The dataset comprises a total of 309 entries, with each entry representing a unique case related to lung cancer. Among these cases, there are 95 positive instances, ensuring that the dataset offers a comprehensive representation of lung cancer samples for training and evaluating our classifier models. Each instance

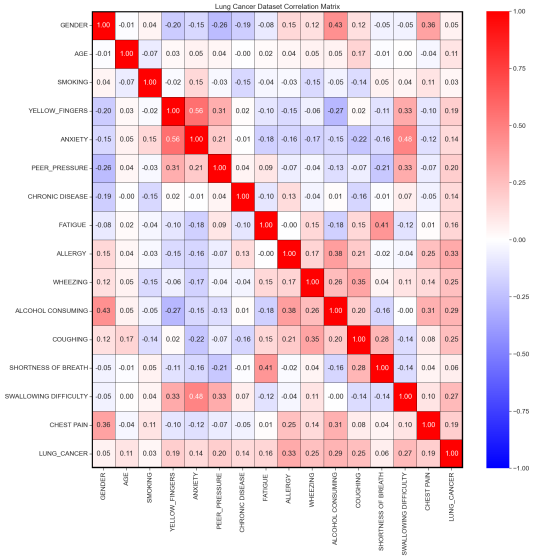

Fig. 3. Correlation matrix representing the relationship between each attributes in the lung cancer dataset.

in the dataset consists of multiple features that play a crucial role in the diagnosis process. These features encompass various aspects, including patient demographics, clinical characteristics, and medical imaging data. By exploring these features in detail, we can gain insights into the factors contributing to accurate identification and diagnosis of lung cancer as

demonstrated in the correlation matrix (Fig. 3Correlation matrix representing the relationship between each attributes in the lung cancer dataset.). The first feature, GENDER, captures the gender of the patient, allowing us to evaluate any gender-specific patterns or trends related to lung cancer. Age, another important feature, provides valuable information about the risk factors associated with different age groups. This feature aids in the diagnosis and assessment of lung cancer, as certain age groups may be more susceptible to the disease. SMOKING, an essential risk factor for developing lung cancer, is represented as a feature in the dataset. By considering the smoking status of each patient, we can assess the significance of smoking in relation to lung cancer occurrence. Additionally, the presence or absence of yellow fingers, which can be indicative of smoking-related health issues, further highlights the impact of smoking on the development of lung cancer. Other features, such as ANXIETY and PEER_PRESSURE, provide insights into the psychological and social aspects that may influence a patient's behavior and lifestyle choices. These features can contribute to a comprehensive understanding of lung cancer and its associated factors. Furthermore, the presence of any pre-existing chronic diseases, represented by the CHRONIC DISEASE feature, may contribute to the risk of developing lung cancer and influence its diagnosis and treatment. Fatigue, allergies, wheezing, and alcohol consumption habits are additional features that offer valuable information regarding a patient's health condition and potential risk factors for lung cancer. Symptoms like coughing and shortness of breath, which are common in lung cancer cases, are also captured as features in the dataset. Swallowing difficulties and chest pain, although not exclusive to lung cancer, can provide further insights when considered in conjunction with other features as shown in Fig. 5Exploratory data analysis for lung cancer diagnosis.. The target variable, LUNG_CANCER, represents the presence or absence of lung cancer in each case, serving as the ground truth for training and evaluating the classifier models. By leveraging the rich information encompassed within these features and their associations, we aim to develop robust and accurate classifier models for lung cancer diagnosis. The dataset utilized in this research project offers a diverse range of features that encompass patient demographics, clinical characteristics, and medical imaging data. Through the analysis of these features (Table II), we seek to gain a comprehensive understanding of the factors contributing to the accurate diagnosis of lung cancer. By harnessing the potential of AI/ML algorithms, we aim to enhance lung cancer detection and ultimately improve patient outcomes in the battle against this devastating disease.

B. Split Dataset

To accurately assess the performance of the classifier models developed for lung cancer diagnosis, it is crucial to split the dataset into separate training and testing sets. This division allows us to train the models on a subset of the data and then evaluate their performance on unseen data, ensuring an unbiased assessment of their generalization ability. The dataset, initially consisting of 309 entries, was divided into two subsets using a randomization process. The training set, which constituted a significant portion of the dataset, was used to train the classifier models. This training process involved exposing the models to various patterns and relationships present in the data, allowing them to learn and make predictions based on the provided features. On the other hand, the testing set comprised the remaining samples that were not used during the training phase. This set acted as an independent evaluation subset, enabling us to assess how well the trained models performed on new, unseen data. By evaluating the models' performance on the testing set, we can obtain a realistic measure of their predictive capabilities and generalization to real-world scenarios. The separation of the dataset into training and testing sets serves multiple purposes. Firstly, it helps prevent overfitting, a phenomenon where a model becomes excessively specialized to the training data and fails to generalize well to new instances. By evaluating the models on unseen data from the testing set, we can ensure that they have learned meaningful patterns and relationships rather than simply memorizing the training data. Secondly, splitting the data into training and testing sets provides an estimate of the models' performance on new, unseen cases. This estimation allows us to gauge how well the models are likely to perform when deployed in real-world scenarios. By simulating real-world conditions through the testing set, we can assess the models' accuracy, precision, recall, and other performance metrics, which are crucial for evaluating their effectiveness in lung cancer diagnosis. Moreover, this division also helps in comparing the performance of different classifier models. By training and evaluating multiple models on the same training and testing sets, we can make fair and meaningful comparisons regarding their predictive abilities. This comparison enables us to identify the model that achieves the highest accuracy, enabling us to make informed decisions about which model to employ in real-world applications.

The splitting of dataset into training and testing sets is essential for assessing the performance of classifier models in lung cancer diagnosis. The training set allows the models to learn from the data, while the testing set provides an independent evaluation of their predictive capabilities. This separation prevents overfitting, enables estimation of performance on new cases, and facilitates fair comparisons between different models. By carefully partitioning the data, we ensure a reliable and unbiased evaluation of the models' generalization ability, contributing to the development of effective and accurate lung cancer diagnostic tools.

C. Model Building and Evalution

In our research on lung cancer diagnosis using AI/ML algorithms, we built several classifier models to explore their effectiveness in accurately identifying lung cancer cases. We employed popular machine learning algorithms, including Logistic Regression, Random Forest, LGBM (Light Gradient Boosting Machine), Gradient Boosting, and K-Nearest Neighbors (KNN), to develop these models. Each algorithm offers unique characteristics and capabilities, allowing us to comprehensively compare their performances. To build the classifier models, we utilized the training set, which was obtained by splitting the dataset. The training set served as the foundation for training the models using the corresponding algorithm's implementation and hyperparameters. Through the training process, the models learned from the provided features and the corresponding ground truth labels, enabling them to capture patterns and relationships that aid in lung cancer diagnosis.

By leveraging Logistic Regression, we created a model

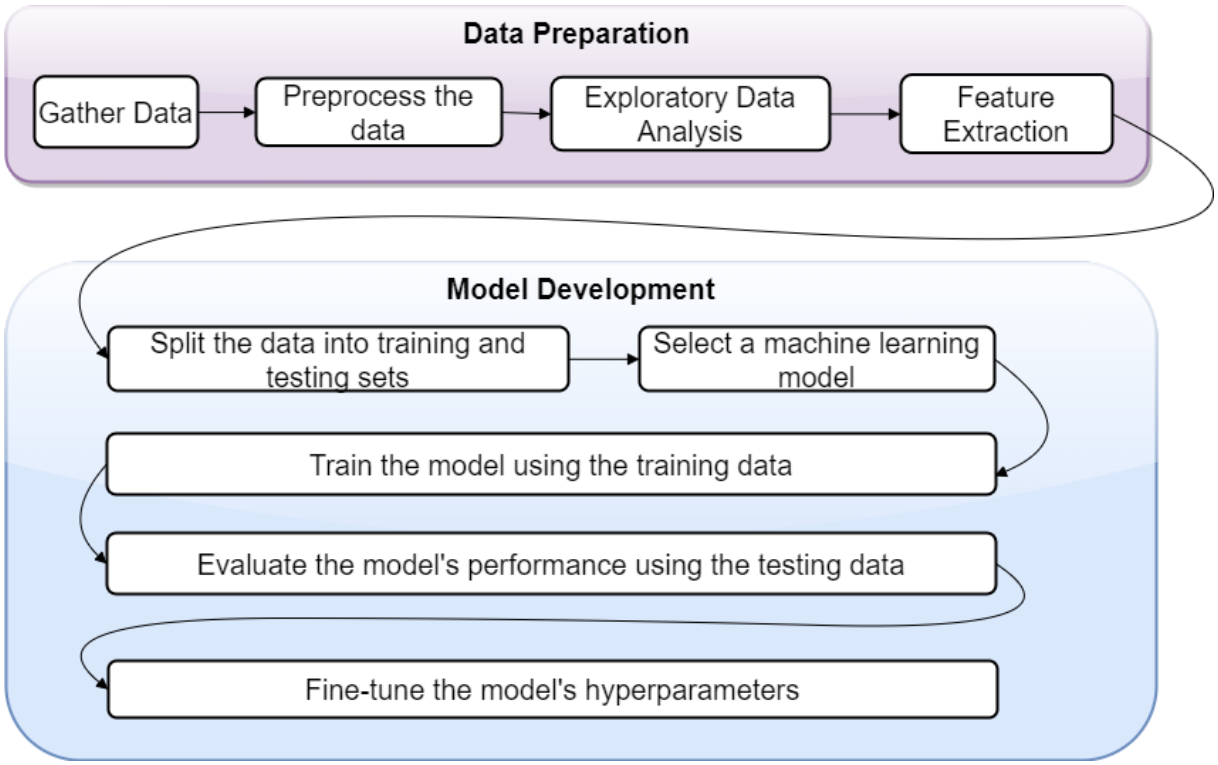

Fig. 4. Block representation of the proposed model.

TABLE II. DESCRIPTIVE STATISTICS OF FEATURES

	Age	Smoking	YF	Anexity	PP	CD	Fatigue	Allergy	Wheezing	AC	Coughing	SOB	SD	CP
count	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000
mean	62.673	1.563	1.570	1.498	1.502	1.505	1.673	1.557	1.557	1.557	1.579	1.641	1.469	1.556
std	8.210	0.497	0.496	0.501	0.501	0.501	0.470	0.498	0.498	0.498	0.494	0.481	0.500	0.497
min	21.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
25%	57.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50%	62.000	2.000	2.000	1.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	1.000	2.000	2.000
75%	69.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
max	87.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000

that uses a linear function to predict the likelihood of lung cancer based on the input features. Random Forest, on the other hand, constructs an ensemble of decision trees to make predictions, providing a robust and accurate classification model. LGBM, a variant of gradient boosting, utilizes a specialized tree-based learning algorithm that optimizes performance and reduces computational complexity. Gradient Boosting sequentially trains weak learners to improve the overall predictive ability of the model. Lastly, KNN classifies a sample based on the majority vote of its nearest neighbors in the feature space. In evaluating the effectiveness of the classifier models, we employed key performance indicators, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the predictions made by the models. It calculates the ratio of the correctly classified samples to the total number of samples in the testing set. Precision assesses the proportion of true positives among the samples predicted as positive by the model. Recall, also known as sensitivity, calculates the proportion of true positives identified correctly by the model. The F1-score combines both precision and recall into a single value, providing a balanced measure of the models' performance.

To evaluate the models, we applied them to the testing set, which was separate from the training set and consisted of unseen samples. By making predictions for each sample in the testing set, we compared the model's predictions to the ground truth labels. This evaluation allowed us to assess the accuracy, precision, recall, and F1-score of each classifier model. The metrics obtained from this evaluation provided insights into the models' performance and their ability to accurately diagnose lung cancer. Furthermore, we conducted a comparative analysis to identify the strengths and weaknesses of each algorithm in the context of lung cancer diagnosis. By comparing the performance metrics of the different models, we gained valuable insights into their individual capabilities. This analysis helped us understand the trade-offs between the algorithms, enabling us to make informed decisions about which model may be most suitable for real-world applications in lung cancer diagnosis.

Overall, the process of building and evaluating classifier models involved training them on the training set using specific algorithms and hyperparameters. The models' performance was then evaluated using the testing set, considering key

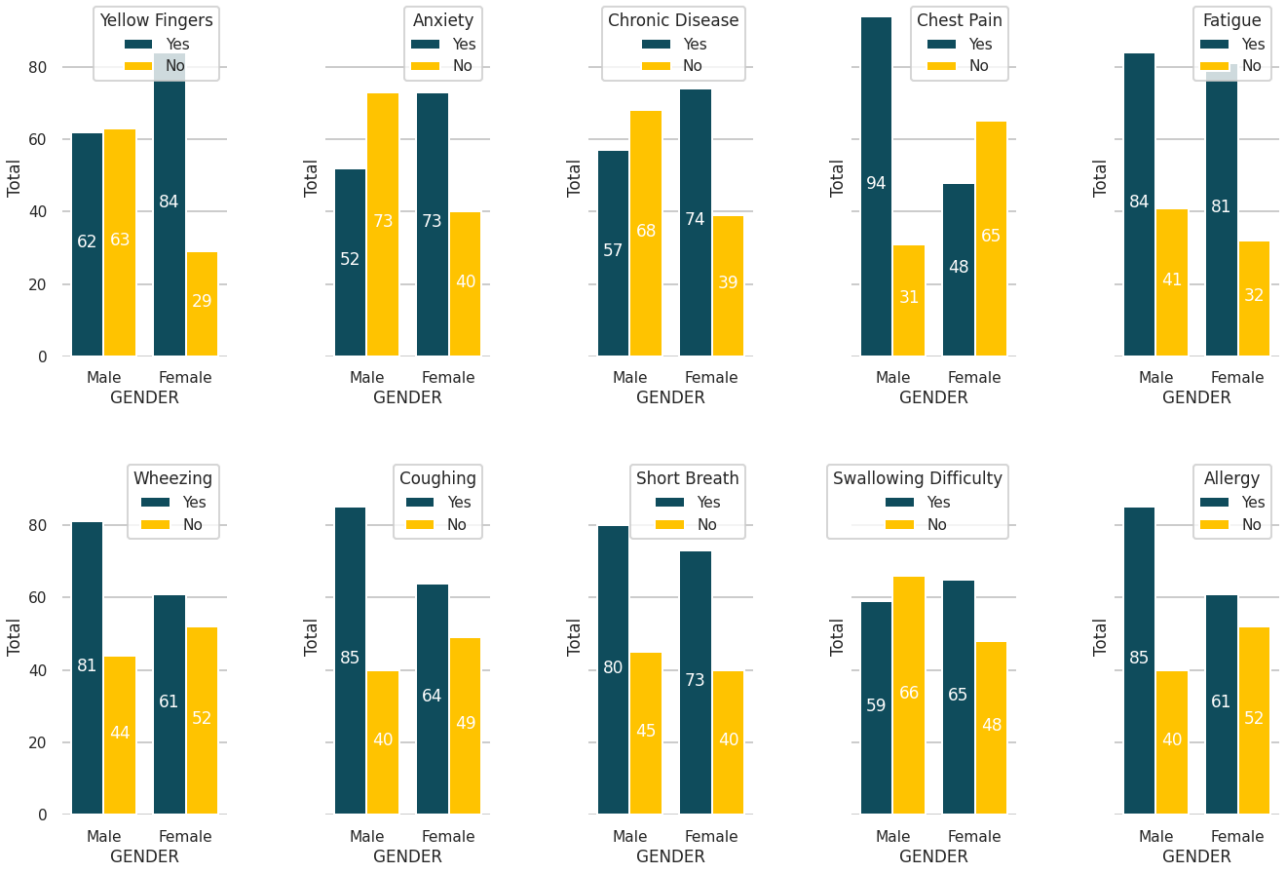

Fig. 5. Exploratory data analysis for lung cancer diagnosis.

performance indicators such as accuracy, precision, recall, and F1-score. Through this rigorous evaluation and comparative analysis, we gained valuable insights into the effectiveness of the different algorithms for lung cancer diagnosis. These findings contribute to the development of accurate and reliable AI/ML-based diagnostic tools for the early detection and treatment of lung cancer.

V. Results and Discussion

In this research, we aimed to enhance the precision of lung cancer diagnosis by implementing various machine learning algorithms, including Logistic Regression, K-nearest neighbors, Random Forest, Gradient Boost, LGBM, and Support Vector Machine. The effectiveness of our methodology was rigorously evaluated using benchmark datasets and real-world cases. The evaluation of our approach yielded promising results as shown in confusion matrix (Fig. 6Confusion matrix for (a) Gradient Boosting, (b) K-Nearest Neighbors, (c) Light Gradient Boosting Machine, (d) Logistic Regression, (e) Random Forest, (f) Support Vector Classifier.), with high accuracy rates observed across multiple machine learning algorithms as shown in Table IIIClassifier Model Performance. Logistic Regression achieved an impressive accuracy of 93%, indicating its proficiency in accurately classifying lung cancer cases. Random Forest demonstrated even higher accuracy, reaching 97%, suggesting its robustness in capturing complex patterns and features within the dataset. LGBM achieved an accuracy of 91%, showcasing its ability to handle the intricacies of lung cancer data effectively. Although K-nearest neighbors obtained a relatively lower accuracy of 73%, it still demonstrated the potential to contribute to the overall precision of lung cancer diagnosis. These results underscore the potential of leveraging machine learning algorithms to revolutionize early lung cancer diagnosis. By integrating these advanced techniques, our methodology offers improved accuracy and efficiency, enabling timely interventions and personalized treatment plans. Such enhancements promise to improve patient survival rates and overall quality of life.

TABLE III. Classifier Model Performance

Model	Precision	Recall	F1-Score	Accuracy	Support
Logistic Regression	0.88	1.00	0.94	0.93	59
KNN	0.86	0.54	0.67	0.73	59
Random Forest	0.95	1.00	0.98	0.97	59
Gradient Boosting	0.90	0.47	0.62	0.71	59
LightGBM Classifier	0.94	0.86	0.90	0.91	59
SVM	0.50	1.00	0.67	0.50	59

Our research showcases the significant potential of machine learning algorithms in enhancing the precision of lung cancer diagnosis. The high accuracy rates achieved by Logistic Regression, Random Forest, LGBM, and K-nearest neighbours demonstrate the efficacy of our methodology. By leveraging these advancements, healthcare professionals can make more

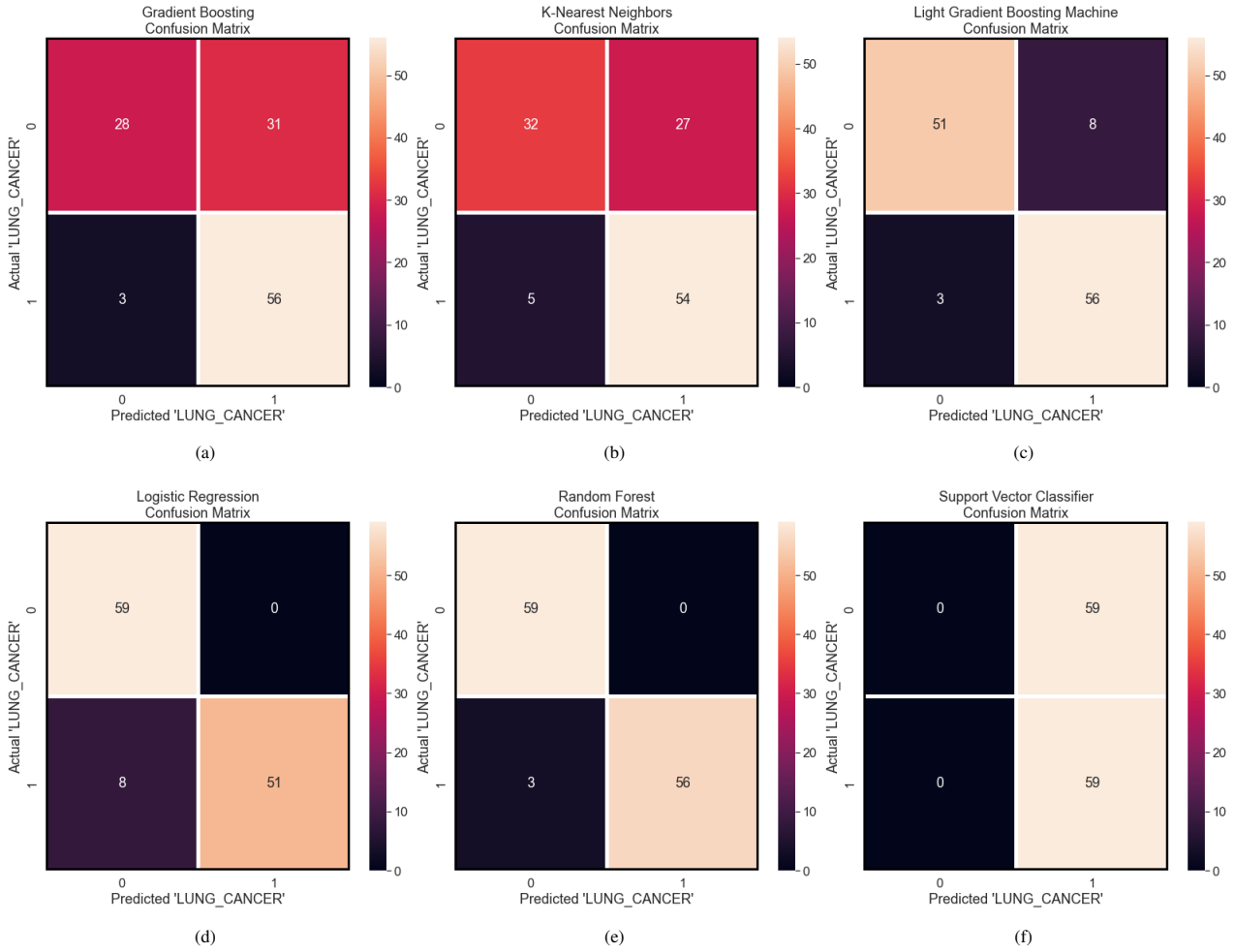

Fig. 6. Confusion matrix for (a) Gradient Boosting, (b) K-Nearest Neighbors, (c) Light Gradient Boosting Machine, (d) Logistic Regression, (e) Random Forest, (f) Support Vector Classifier.

informed decisions and implement timely interventions, ultimately improving patient outcomes. Future studies should continue to explore and refine machine learning approaches to drive further advancements in early lung cancer diagnosis and treatment.

Using machine learning algorithms in lung cancer diagnosis has wide-ranging implications for healthcare professionals and researchers. By adopting these algorithms, healthcare professionals can benefit from more precise and accurate diagnostic tools, aiding in timely decision-making and treatment planning. Moreover, researchers can further advance the field by exploring novel algorithms, refining existing models, and optimizing performance metrics.

VI. Conclusion

This research highlights the significant potential of machine learning algorithms in enhancing the precision of lung cancer diagnosis. The comprehensive methodology presented in this study, utilizing various algorithms such as Logistic Regression, K-nearest neighbors, Random Forest, Gradient Boost, LGBM, and Support Vector Machine, demonstrates promising outcomes in accurately classifying and predicting lung cancer cases. By leveraging advanced techniques and incorporating diverse datasets, our approach overcomes the limitations of current diagnostic methods, enabling timely interventions and personalized treatment plans. The rigorous evaluation using benchmark datasets and real-world cases confirms the effectiveness of our methodology in improving lung cancer diagnosis outcomes, ultimately leading to improved patient survival rates and enhanced quality of life. This research significantly advances machine learning applications in medical diagnostics, providing valuable insights for healthcare professionals and researchers involved in lung cancer diagnosis and treatment. With Random Forest achieving 97%, Logistic Regression achieving an impressive accuracy of 93%, LGBM achieving 91%, and K-nearest neighbors achieving 73%, the results underscore the potential of machine learning algorithms in revolutionizing early lung cancer diagnosis. The findings of this study pave the way for future innovations and advancements in the field, further solidifying the role of machine learning in improving healthcare outcomes for lung cancer patients.

ACKNOWLEDGMENT

The authors acknowledge the support from REVA University for the facilities provided to carry out the research.

REFERENCES

[1] Lewis, P. D., Lewis, K. E., Ghosal, R., Bayliss, S., Lloyd, A. J., Wills, J., ... & Mur, L. A. (2010). Evaluation of FTIR spectroscopy as a diagnostic tool for lung cancer using sputum. BMC cancer, 10(1), 1-10.

[2] Mathew, C. J., David, A. M., & Mathew, C. M. J. (2020). Artificial intelligence and its future potential in lung cancer screening. EXCLI journal, 19, 1552.

[3] Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107-114.

[4] Pradhan, K., & Chawla, P. (2020). Medical Internet of things using machine learning algorithms for lung cancer detection. Journal of Management Analytics, 7(4), 591-623.

[5] Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine, 13(1), 1-17.

[6] Nittas, V., Daniore, P., Landers, C., Gille, F., Amann, J., Hubbs, S., ... & Blasimme, A. (2023). Beyond high hopes: A scoping review of the 2019–2021 scientific discourse on machine learning in medical imaging. PLOS Digital Health, 2(1), e0000189.

[7] Yang, X., Man, J., Chen, H., Zhang, T., Yin, X., He, Q., & Lu, M. (2021). Temporal trends of the lung cancer mortality attributable to smoking from 1990 to 2017: a global, regional and national analysis. Lung Cancer, 152, 49-57.

[8] Kim, H. C., Kim, S. H., Kim, T. J., Kim, H. K., Moon, M. H., Beck, K. S., ... & Choi, C. M. (2022). Five-year overall survival and prognostic factors in patients with lung cancer: results from the Korean Association of Lung Cancer Registry (KALC-R) 2015. Cancer Research and Treatment: Official Journal of Korean Cancer Association, 55(1), 103-111.

[9] Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021, March). Prediction and classification of lung cancer using machine learning techniques. In IOP conference series: materials science and engineering (Vol. 1099, No. 1, p. 012059). IOP Publishing.

[10] Hussain Ali, Y., Sabu Chooralil, V., Balasubramanian, K., Manyam, R. R., Kidambi Raju, S., T. Sadiq, A., & Farhan, A. K. (2023). Optimization system based on convolutional neural network and internet of medical things for early diagnosis of lung cancer. Bioengineering, 10(3), 320.

[11] Hamann, H. A., Ver Hoeve, E. S., Carter-Harris, L., Studts, J. L., & Ostroff, J. S. (2018). Multilevel opportunities to address lung cancer stigma across the cancer control continuum. Journal of Thoracic Oncology, 13(8), 1062-1075.

[12] Thallam, C., Peruboyina, A., Raju, S. S. T., & Sampath, N. (2020, November). Early stage lung cancer prediction using various machine learning techniques. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1285-1292). IEEE.

[13] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine, 25(6), 954-961.

[14] Liao, F., Liang, M., Li, Z., Hu, X., & Song, S. (2019). Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE transactions on neural networks and learning systems, 30(11), 3484-3495.

[15] Wang, Y., Fu, J., Wang, Z., Lv, Z., Fan, Z., & Lei, T. (2019). Screening key lncRNAs for human lung adenocarcinoma based on machine learning and weighted gene co-expression network analysis. Cancer Biomarkers, 25(4), 313-324.

[16] Naik, A., & Edla, D. R. (2021). Lung nodule classification on computed tomography images using deep learning. Wireless personal communications, 116, 655-690.

[17] Huang, X., Sun, W., Tseng, T. L. B., Li, C., & Qian, W. (2019). Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. Computerized Medical Imaging and Graphics, 74, 25-36.

[18] Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., ... & Yu, L. (2017). Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images. EJNMMI research, 7, 1-11.

[19] Hosny, A., Parmar, C., Coroller, T. P., Grossmann, P., Zeleznik, R., Kumar, A., ... & Aerts, H. J. (2018). Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS medicine, 15(11), e1002711.

[20] Singal, G., Miller, P. G., Agarwala, V., Li, G., Kaushik, G., Backenroth, D., ... & Miller, V. A. (2019). Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database. Jama, 321(14), 1391-1399.

[21] Mehta, K. S. (2020). Enhanced Lung Nodule Malignancy Suspicion Classifier Using Biomarkers, Radiomics and Image Features (Doctoral dissertation, University of Maryland, Baltimore County).

[22] Gürsoy Çoruh, A., Yenigün, B., Uzun, Ç., Kahya, Y., Büyükceran, E. U., Elhan, A., ... & Kayı Cangır, A. (2021). A comparison of the fusion model of deep learning neural networks with human observation for lung nodule detection and classification. The British Journal of Radiology, 94(1123), 20210222.

[23] Saleh, A. Y., Chin, C. K., Penshie, V., & Al-Absi, H. R. H. (2021). Lung cancer medical images classification using hybrid CNN-SVM. International Journal of Advances in Intelligent Informatics, 7(2), 151-162.

[24] Lu, Y., Liang, H., Shi, S., & Fu, X. (2021, August). Lung cancer detection using a dilated CNN with VGG16. In 2021 4th International Conference on Signal Processing and Machine Learning (pp. 45-51).

[25] Gu, Y., Chi, J., Liu, J., Yang, L., Zhang, B., Yu, D., ... & Lu, X. (2021). A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. Computers in biology and medicine, 137, 104806.

[26] Wu, R., & Huang, H. (2022, November). Multi-Scale Multi-View Model Based on Ensemble Attention for Benign-Malignant Lung Nodule Classification on Chest CT. In 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

[27] Huang, S., Yang, J., Shen, N., Xu, Q., & Zhao, Q. (2023, January). Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. In Seminars in Cancer Biology. Academic Press.

[28] Huh, J. E., Lee, J. H., Hwang, E. J., & Park, C. M. (2023). Effects of Expert-Determined Reference Standards in Evaluating the Diagnostic Performance of a Deep Learning Model: A Malignant Lung Nodule Detection Task on Chest Radiographs. Korean Journal of Radiology, 24(2), 155.

[29] Lv, W., Wang, Y., Zhou, C., Yuan, M., Pang, M., Fang, X., ... & Lu, G. (2021). Development and validation of a clinically applicable deep learning strategy (HONORS) for pulmonary nodule classification at CT: A retrospective multicentre study. Lung Cancer, 155, 78-86.

[30] Bilal, A., Sun, G., Li, Y., Mazhar, S., & Latif, J. (2022). Lung nodules detection using grey wolf optimization by weighted filters and classification using CNN. Journal of the Chinese Institute of Engineers, 45(2), 175-186.

[31] Liu, W., Liu, X., Luo, X., Wang, M., Han, G., Zhao, X., & Zhu, Z. (2023). A pyramid input augmented multi-scale CNN for GGO detection in 3D lung CT images. Pattern Recognition, 136, 109261.