International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 1, January 2025)

Review of Lung Cancer Diseases Prediction using AI Techniques

[1]Anand Kishor Agnihotri, [2]Manish Gupta, [3]Dr Anshuj Jain

[1]Research Scholar, Dept. of Electronics and Communication Engineering, SCOPE College of Engineering, Bhopal, India,
[2]Assistant Professor, Dept. of Electronics and Communication Engineering, SCOPE College of Engineering, Bhopal, India
[3]Associate Professor & HOD, Dept. of Electronics and Communication Engineering, SCOPE College of Engineering, Bhopal, India

***Abstract*— Lung cancer remains one of the most fatal diseases worldwide, accounting for a significant proportion of cancer-related deaths annually. Early detection and accurate prediction are critical in improving survival rates and enabling timely interventions. With the advent of Artificial Intelligence (AI), particularly machine learning (ML) and deep learning (DL) techniques, the healthcare industry has witnessed transformative advancements in the diagnosis and prediction of lung cancer. AI-powered methods leverage vast datasets, including medical imaging, genetic profiles, and clinical records, to uncover patterns that are often imperceptible to human analysis. This review explores the state-of-the-art AI techniques employed for lung cancer prediction, focusing on their methodologies, applications, and effectiveness.**

***Keywords*— *Machine Learning, Lung, Classification, Diagnosis* .**

I. INTRODUCTION

Lung cancer is a leading cause of morbidity and mortality worldwide, accounting for approximately 2.2 million new cases and 1.8 million deaths annually, as reported by the World Health Organization (WHO). Despite advancements in medical science, the prognosis for lung cancer remains grim, primarily due to late-stage diagnosis in the majority of patients. Early detection is pivotal in improving survival rates, as the chances of effective treatment significantly decline in advanced stages of the disease. However, traditional diagnostic methods such as imaging, biopsy, and clinical assessment often face limitations in accuracy, speed, and accessibility. The need for innovative approaches to address these challenges has propelled the integration of Artificial Intelligence (AI) into lung cancer research.

AI has emerged as a transformative tool in modern healthcare, offering unparalleled capabilities in data processing, pattern recognition, and decision-making. By leveraging large-scale datasets, AI techniques can uncover complex relationships between variables, enabling precise and early prediction of diseases. In the context of lung cancer, AI-powered systems are being utilized to analyze medical imaging data, such as chest X-rays and computed tomography (CT) scans, as well as genomic and proteomic profiles. These systems aim to enhance the sensitivity and specificity of lung cancer detection, reducing the burden on healthcare systems and improving patient outcomes.

Figure 1: Lungs Sample (google image)

The application of machine learning (ML) and deep learning (DL) has been particularly noteworthy in this field. ML algorithms, such as Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (kNN), have demonstrated significant potential in classifying cancerous and non-cancerous samples. On the other hand, DL techniques, especially Convolutional Neural Networks (CNNs), have revolutionized image-based diagnostics by achieving human-level accuracy in tasks such as tumor localization and

classification. Furthermore, ensemble learning models that combine multiple algorithms are being developed to optimize prediction performance.

Despite these promising advancements, several challenges persist in the application of AI for lung cancer prediction. The quality and diversity of data are critical factors that influence the performance of AI models. Issues such as data imbalance, missing values, and the lack of standardized datasets hinder the generalizability of these models. Additionally, interpretability and transparency remain significant concerns, as the "black box" nature of many AI systems limits their acceptance in clinical practice. Ethical considerations, including patient privacy and algorithmic bias, further complicate the integration of AI into healthcare.

This review delves into the cutting-edge AI techniques used for lung cancer prediction, emphasizing their methodologies, applications, and limitations. It aims to provide a comprehensive understanding of how AI is shaping the future of lung cancer diagnosis and highlight the pathways for future research and development in this domain. By bridging the gap between technology and medicine, this study underscores the transformative potential of AI in combating one of the deadliest diseases of our time.

II. LITERATURE SURVEY

J. S. S. Araújo et al.,[1] The features extraction on thorax computerised tomography images is performed using the radiological densities of human tissues in Hounsfield Units. In order to evaluate the efficacy of the proposed method in conjunction with four machine learning classifiers, we conducted a comparison with the Grey Level Co-occurrence Matrix and Statistical Moments. In general, the findings indicated that the proposal obtained the highest accuracy ratios and the shortest duration among all experiments conducted. Consequently, we regard our proposal as a viable alternative that can be implemented in real-time applications.

X. Qian et al., [2] in their presentation of lung diseases in Healthy, H1N1, and CAP cases. We conduct extensive experiments on a chest CT imaging dataset with a total of 734 patients, including 251 healthy individuals, 245 patients with lung diseases, 105 H1N1 patients, and 133 CAP patients, in order to further demonstrate the efficacy of our model. The quantitative results, which include a plethora of metrics, demonstrate the superiority of our proposed model in both slice- and patient-level classification tasks. Even more significantly, the lesion location maps that are generated render our system more comprehensible and valuable to clinicians.

T. K. K. Ho et al.,[3] is to either convert knowledge from ponderous teacher models into lightweight student models or to self-train these student models in order to produce weakly supervised multi-label lung disease classifications. The visualisations employed in saliency maps of the pathological regions where an abnormality was located were supported by multi-task deep learning architectures, in addition to multi-class classification, which was the foundation of our approach. A self-training KD framework, in which the model learnt from itself, was demonstrated to outperform both the well-established baseline training procedure and the conventional KD, attaining AUC improvements of up to 6.39% and 3.89%, respectively. In comparison to the current deep learning baselines, our approach effectively surmounted the interdependency of 14 inadequately annotated thorax maladies and facilitated the state-of-the-art classification through application to the publicly available ChestX-ray14 dataset.

AFS-DF on the lungs diseases -19 dataset is presented by L. Sun et al., [4]. The dataset includes 1495 patients with lungs diseases -19 and 1027 patients with community-acquired pneumonia (CAP). Our method has demonstrated accuracy (ACC), sensitivity (SEN), specificity (SPE), AUC, precision, and F1-score of 91.79%, 93.05%, 89.95%, 96.35%, 93.10%, and 93.07%, respectively. The experimental results on the lungs diseases dataset indicate that the proposed AFS-DF outperforms four widely used machine learning methods in the classification of lungs diseases vs. CAP.

C. Baloescu et al.,[5] In order to develop and evaluate the deep learning algorithm based on deep convolutional neural networks, we employed ultrasound recordings (n = 400) from an existing database of ED patients to serve as training and test sets. Expert human interpretations of binary and severity classifications (a scale of 0-4) were contrasted with the algorithmic interpretations of the images. In comparison to an expert read, our model demonstrated a sensitivity of 93% (95% confidence interval (CI) 81%-98%) and a specificity of 96% (95% CI 84%-99%) for the presence or absence of B-lines. The kappa ratio was 0.88 (95% CI 0.79-0.97). A weighted kappa of 0.65 (95% CI 0.56- 074) was obtained for the severity classification model to expert agreement.

J. X. Wu et al.,[6] Subjects with typical lung diseases are screened using a multilayer machine vision classifier that incorporates a radial Bayesian network and grey relational analysis. The NIH chest X-ray database (NIH Clinical Centre)

is utilised to enrol anterior-posterior chest X-ray images. The proposed multilayer machine vision classifier is employed to assist in the diagnosis of common lung diseases on specific bounding ROIs using digital chest X-ray images. The performance of the proposed multilayer classifier for the rapid screening of lung lesions on digital chest X-ray images is evaluated using mean recall (%), mean precision (%), mean accuracy (%), and mean F1 score (0.8981), respectively, with K-fold cross-validation.

S. Roy et al., [7] presented a novel approach to the video-level aggregation of frame scores that is based on uniforms. Lastly, we evaluate the performance of cutting-edge deep models in the estimation of pixel-level segmentations of lung disease imaging biomarkers. The proposed dataset has yielded satisfactory results in all of the tasks that were examined, which will facilitate future research on deep learning for the assisted diagnosis of pulmonary diseases from LUS data.

S. Pang et al., [8] suggest a deep learning model for the identification of lung cancer type from CT images in patients at Shandong Provincial Hospital. It faces a dual challenge: the limited number of patient data acquired and the inadequacy of artificial intelligent models trained on public datasets in meeting these practical requirements. The two-fold problem is resolved by employing image rotation, translation, and transformation methods to expand and balance our training data. Subsequently, densely connected convolutional networks (DenseNet) are employed to classify malignant tumours from images collected. Finally, the adaptive boosting (adaboost) algorithm is employed to aggregate multiple classification results in order to enhance classification performance.

H. Yazdani et al., [9] introduce a method that assesses the migration of samples from one cluster to another. This method enables us to identify critical samples in advance that have the potential to be a part of other clusters in the near future. In a lung cancer case-control investigation, BFPM was implemented to analyse the metabolomics of individuals. Metabolomics may function as robust biomarkers of the current disease process by providing proximate molecular signals to the actual disease processes. The objective is to determine whether it is possible to distinguish between the serum metabolites of a healthy individual and those of a person with lung cancer. BFPM was employed to identify critical samples, evaluate pathology data, and observe certain discrepancies.

F. Yan et al.,[10] The chest X-ray is a straightforward and cost-effective medical tool for auxiliary diagnosis, and as a result, it has become a standard component of physical examinations for physicians. By utilising deep learning techniques, we investigate the abnormality classification problem of chest X-rays using 40167 images of chest radiographs and corresponding reports. Initially, we suggest an annotation method that is based on the anomalous portion of the images, as radiology reports are typically templated by the aberrant physical regions. Secondly, we utilise the long short-term memory (LSTM) model to automatically annotate the remaining unlabelled data, building on a limited number of reports that are manually annotated by professional radiologists. The precision value, recall value, and F1-score all exceed 0.88 in the accurate annotation of images.

A. Rao et al.,[11] The accumulation of excessive air and water in the lungs results in the impairment of respiratory function and is a prevalent cause of patient hospitalisation. Physicians can evaluate patients' respiratory conditions by employing non-invasive and compact methods to detect changes in lung fluid accumulation. In this study, a digital stethoscope instrument and an acoustic transducer are suggested as a targeted solution to address this clinical requirement. Measurable changes in the structure of the lungs can be employed to evaluate lung pathology. We standardise this procedure by transmitting a controlled signal through the airways of six healthy subjects and six patients with lung disease. Mel-frequency cepstral coefficients and spectroid audio features, which are frequently employed in classification for music retrieval, are extracted to differentiate between healthy and diseased subjects. We exhibit a 91.7% accuracy in the differentiation between healthy subjects and patients with lung pathology by employing the K-nearest neighbours algorithm.

O. P. Singh et al.,[12] At present, the clinic employs capnography to measure carbon dioxide (CO 2) waveforms in order to estimate respiratory rate and end-tidal CO2 (EtCO 2). Nevertheless, the asthmatic condition is significantly influenced by the morphology of the CO 2 signal. Previous research has demonstrated a robust correlation between a variety of features that quantitatively characterise the shape of the CO 2 signal and are employed to differentiate asthma from non-asthma using pulmonary function tests. However, no reliable progress has been made, and no translation into clinical practice has been achieved. Consequently, this study presents a signal processing algorithm that is relatively straightforward and can be used to automatically differentiate between asthma and non-asthma.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 1, January 2025)

III. CHALLENGES

1. Data-Related Challenges

AI models rely heavily on high-quality, diverse, and extensive datasets to deliver accurate and reliable predictions. However, the healthcare industry faces several data-related hurdles:

- **Data Imbalance:** Medical datasets often exhibit an imbalance where healthy cases significantly outnumber positive cases for lung cancer. This can lead to biased predictions and poor sensitivity to minority classes.
- **Heterogeneous Data Sources:** Combining imaging data (e.g., CT scans, X-rays), clinical records, and genomic data introduces inconsistencies in format, quality, and completeness. Integrating such heterogeneous data is a significant challenge.
- **Limited Access to Data:** Patient privacy regulations, such as HIPAA and GDPR, often limit access to large-scale datasets, restricting the ability to train robust AI models.
- **Anonymization and Labeling:** Ensuring proper anonymization while maintaining data utility and obtaining accurate, expert-labeled datasets is resource-intensive and time-consuming.

2. Model Performance and Generalizability

- **Overfitting to Training Data:** AI models trained on limited or region-specific datasets may perform well in controlled environments but fail to generalize to broader populations.
- **Variability in Medical Imaging:** Differences in imaging protocols, equipment, and resolutions across hospitals can reduce model performance when deployed in diverse clinical settings.
- **Handling Rare Subtypes:** Lung cancer includes various subtypes (e.g., small cell and non-small cell lung cancer), some of which are rare and challenging for AI models to predict accurately due to insufficient training examples.

3. Interpretability and Explainability

Many AI techniques, particularly deep learning models like Convolutional Neural Networks (CNNs), operate as "black boxes," making it difficult for clinicians to understand how predictions are made.

- **Lack of Explainability:** The inability to explain AI decisions undermines trust and acceptance among clinicians. For instance, a false positive without a clear rationale can lead to unnecessary biopsies or emotional distress for patients.
- **Clinical Validation:** Without explainable outcomes, it becomes challenging to validate AI predictions against clinical expertise and established diagnostic protocols.

4. Ethical and Regulatory Concerns

AI in healthcare introduces several ethical and regulatory challenges that need careful consideration:

- **Patient Privacy:** Ensuring data security and maintaining patient confidentiality are critical concerns, especially when dealing with sensitive medical records.
- **Algorithmic Bias:** Biases in training data, such as underrepresentation of specific demographic groups, can lead to discriminatory outcomes.
- **Regulatory Approval:** Gaining regulatory approvals for AI-based diagnostic tools involves stringent and often prolonged validation processes, slowing down their adoption.

IV. CONCLUSION

The application of Artificial Intelligence in lung cancer prediction has demonstrated transformative potential by enabling early detection, enhancing diagnostic accuracy, and improving patient outcomes. Techniques such as machine learning and deep learning have revolutionized the analysis of complex medical datasets, offering unprecedented insights into patterns that aid in precise prediction. However, challenges related to data quality, model generalizability, interpretability, ethical considerations, and clinical integration remain significant barriers to widespread adoption. Addressing these issues through interdisciplinary collaboration, improved

regulatory frameworks, and advancements in AI technology will pave the way for its seamless integration into clinical workflows, ultimately contributing to the fight against one of the deadliest diseases worldwide.

REFERENCES

1. S. S. Araújo Alves, E. de Souza Rebouças, S. A. Freitas de Oliveira, A. Magalhães Braga and P. P. Rebouças Filho, "Lung Diseases Classification by Analysis of Lung Tissue Densities," in IEEE Latin America Transactions, vol. 18, no. 09, pp. 1329-1336, September 2020, doi: 10.1109/TLA.2020.9381790.
2. X. Qian et al., "M^3$Lung-Sys: A Deep Learning System for Multi-Class Lung Pneumonia Screening From CT Imaging," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3539-3550, Dec. 2020, doi: 10.1109/JBHI.2020.3030853.
3. T. K. K. Ho and J. Gwak, "Utilizing Knowledge Distillation in Deep Learning for Classification of Chest X-Ray Abnormalities," in IEEE Access, vol. 8, pp. 160749-160761, 2020, doi: 10.1109/ACCESS.2020.3020802.
4. L. Sun et al., "Adaptive Feature Selection Guided Deep Forest for lungs diseases Classification With Chest CT," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2798-2805, Oct. 2020, doi: 10.1109/JBHI.2020.3019505.
5. C. Baloescu et al., "Automated Lung Ultrasound B-Line Assessment Using a Deep Learning Algorithm," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, no. 11, pp. 2312-2320, Nov. 2020, doi: 10.1109/TUFFC.2020.3002249.
6. J. -X. Wu, P. -Y. Chen, C. -M. Li, Y. -C. Kuo, N. -S. Pai and C. -H. Lin, "Multilayer Fractional-Order Machine Vision Classifier for Rapid Typical Lung Diseases Screening on Digital Chest X-Ray Images," in IEEE Access, vol. 8, pp. 105886-105902, 2020, doi: 10.1109/ACCESS.2020.3000186.
7. S. Roy et al., "Deep Learning for Classification and Localization of lungs diseases -19 Markers in Point-of-Care Lung Ultrasound," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2676-2687, Aug. 2020, doi: 10.1109/TMI.2020.2994459.
8. S. Pang, Y. Zhang, M. Ding, X. Wang and X. Xie, "A Deep Model for Lung Cancer Type Identification by Densely Connected Convolutional Networks and Adaptive Boosting," in IEEE Access, vol. 8, pp. 4799-4805, 2020, doi: 10.1109/ACCESS.2019.2962862.
9. H. Yazdani, L. L. Cheng, D. C. Christiani and A. Yazdani, "Bounded Fuzzy Possibilistic Method Reveals Information about Lung Cancer through Analysis of Metabolomics," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 2, pp. 526-535, 1 March-April 2020, doi: 10.1109/TCBB.2018.2869757.
10. F. Yan, X. Huang, Y. Yao, M. Lu and M. Li, "Combining LSTM and DenseNet for Automatic Annotation and Classification of Chest X-Ray Images," in IEEE Access, vol. 7, pp. 74181-74189, 2019, doi: 10.1109/ACCESS.2019.2920397.
11. A. Rao, S. Chu, N. Batlivala, S. Zetumer and S. Roy, "Improved Detection of Lung Fluid With Standardized Acoustic Stimulation of the Chest," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1-7, 2018, Art no. 3200107, doi: 10.1109/JTEHM.2018.2863366.
12. O. P. Singh, R. Palaniappan and M. Malarvili, "Automatic Quantitative Analysis of Human Respired Carbon Dioxide Waveform for Asthma and Non-Asthma Classification Using Support Vector Machine," in IEEE Access, vol. 6, pp. 55245-55256, 2018, doi: 10.1109/ACCESS.2018.2871091.