iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: III Month of publication: March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67371

A Real-Time Application for Early-Stage Prediction of Lung Cancer and Its Severity Using Machine Learning

Madan Kumar R[1], Manoj S[2], Hemanth L M[3], Manish A R[4], Dr. Gowrishankar B S[5]
[1, 2, 3, 4]*Department of Information Science,* Vidyavardhaka College of Engineering, Mysuru,India
[5]*Assistant Professor, Vidyavardhaka College of Engineering*

Abstract: Lung cancer remains one of the leading causes of cancer deaths around the world. It is known that the chances of a patient's survival dramatically improve with early diagnosis; however, such diagnosis is often inaccurate with traditional diagnostic methods. The contribution of this study is to propose a machine learning based framework for early diagnosis of lung cancer utilizing advanced algorithms including Convolutional Neural Networks (CNNs), ResNet50 and efficient feature extraction techniques. Our model achieved 98 percent accuracy and 97 percent F1 score, outperforming many state-of-the-art techniques using multimodal imaging datasets. Future studies will focus on hybrid models which incorporate nanotechnology to enhance non-invasive diagnostics, developing solutions for scalability challenges.

Index Terms: F1-Score, nanotechnology, CNN

I. INTRODUCTION

Lung cancer, for instance, remains the most aggressive in terms of the annual number of deaths occurring, at around 1.8 million in 2020. Though early detection tends to increase the survival rates tremendously, conventional methods have specific limitations. Techniques such as CT scans and biopsies, although useful, are expensive, invasive, subjective, and have the tendency of producing too many false positives that cause undue delay in treatment. Everything is about to change with the advent of machine learning, removing ambiguity, and delivering accurate predictions both fast and cheaply. This particular study aims at overcoming these challenges through the application of artificial intelligence techniques, with a focus on deep learning methods. The scope of the research is, therefore, to provide a cost-effective and efficient model for the early detection of lung cancer by incorporating data preprocessing, feature extraction, and efficient CNN architectures. Further, the strategy looks to fill gaps in the risk stratification process that will ultimately help in tailoring treatment strategies among patients and thereby enhance patient involvement.

II. BACKGROUND AND SIGNIFICANCE

It is well established that the diagnosis of lung cancer is very challenging due to the insidious onset when the disease is still asymptomatic, and also on the reliance of X-ray, CT scan, and biopsy techniques for diagnosis. Though these are good for later stages, they do not give an adequate overview of tumors during early stages and hence the diagnosis comes too late which results in poor prognosis.

Worldwide statistics of deaths caused by lung cancer also indicate the necessity for better diagnostic methods. Some researchers have pointed out that if the disease is diagnosed in early stages, the five years survival figure can be raised from 15 to more than 50 percent. But it is also worth mentioning that these diagnostic methods have their drawbacks which are subjective judgment bias, cost and case in low resource areas where accessibility is a problem.

Machine learning especially deep learning has held out hope in answering these challenges because it helps in image analysis which reduces human errors. Other approaches to classify lung cancers with the CNNs and LeNet have gained much attention, as being more accurate than traditional practices on architectures such as ResNet50 and others; however problems such as being biased, having a disproportionate representation of a dataset, it is liable to over fit and too much computational power are preventing it from practical usages. New feature extraction techniques such as DOST, histogram equalization, and many more, significantly improved the capability of the CT and PET imaging models in identifying complex features, and their patterns. Further, collection gaps in the datasets were found, and augmentation, like adding noise or rotation of images, does work, so the models will be stronger and more generalizable.

The innovations in this paper have assisted in making hybrid architectures more effective by the introduction of preprocessing, increasing the application of today's lung cancer diagnostics: making them fast, scalably, and accurate; it also provides solutions with integration multimodal data within study designs, presenting ways to improving interpretability and clinical validity.

III. RELATED WORK

During the last decade, machine learning techniques have gained more usage in the detection of lung cancer because of their processing capabilities and the ability to extract meaningful patterns from large datasets. There are several approaches with improvements and drawbacks.

Tasnim et al. (2024) attempted using CNNs and ResNet50 for the classification of CT scans. It reached a very high accuracy since these two techniques can have more depth in processing the spatial hierarchies. However, the work pointed out that there was a requirement of strong preprocessing of data for solving imbalances in real-world datasets. Similarly, Shafiq et al. (2022) have utilized GoogLeNet combined with feature selection based on fuzzy logic which increases earlystage Small Cell Lung Cancer detection rates. However, this model necessitates enormous computational power that has constrained its scaling.

Other scientists have made use of multimodal imaging techniques to enhance the diagnosis accuracy. Wahengbam et al. (2023) fused CT and PET scan images in order to utilize the complementarities of the information and hence enhancing specificity and sensitivity for detecting lung cancer. This is indeed innovative multimodal imaging technique but led to complexity in integrating and processing data.

Applications of CapsNet have also been explored in medical imaging. The spatial relationship-preserving property made CapsNet a potential tool for the detection of lung cancer, as recent studies have revealed. However, the method failed with large datasets because it is computationally inefficient.

It is further built upon such efforts by including the advanced architectures, which are CNNs and ResNet50, along with the feature extraction techniques, namely DOST and histogram equalization. The proposed framework is in contrast to all the previous methods, because it focuses on realtime applicability through overcoming challenges with dataset diversity, computational requirements, and interpretability.

IV. RESEARCH METHOD

A. Data Collection

Medical Imaging Data: CT and PET scans from medical institutions are collected in order to have high-resolution images to analyze. Such images will be the source of data for training and validation of the machine learning models. Biomarker Data: Blood samples and relevant biomarkers like miR-155, are collected through collaboration with clinical partners that facilitate the use of non-invasive diagnosis methods for early detection of lung cancer.

B. Model Development

A multi-layered deep learning architecture is used to analyze imaging data: Convolutional Neural Networks (CNNs): The architecture uses CNNs to extract features from CT and PET scans, capturing complex patterns associated with lung nodules and tumor characteristics. Hybrid Models: Integrating SVMs and other machine learning algorithms along with the attention mechanism allows the model to enhance its ability in malignancy prediction and also its assessment on tumor invasiveness by features extracted.

C. Algorithm Implementation

Deep Learning Training: The CNNs get trained with a combination of augmented datasets and labeled imaging data which helps improve the model's accuracy and robustness in detecting early-stage lung cancer. Predictive Analytics: Machine learning algorithms to correlate imaging features with clinical outcomes for real-time predictions of lung cancer severity and stratification of risk for appropriate treatment.

V. LIMITATIONS AND RISKS

Despite promising applications toward achieving enhanced diagnostic accuracy in early stages of lung cancer, its development and usage come with inherent limitations and risks, which need to be admitted. Technical, operational, and clinical factors can act against the effectiveness and scalability and feasibility of such a system.

1) *Data Dependency:* The performance of this system relies heavily on the presence of good-quality imaging data and biomarker information. Variability in imaging quality, patient demographics, and methods of sample collection may introduce errors into the model regarding accuracy and generalizability.

2) *Clinical Validation Challenges:* The integration of machine learning models into clinical practice demands an extensive validation process in the form of clinical trials. It is possible that these models may not perform in diverse real-world settings and thus may lead to misdiagnosis or delayed treatment.
3) *Regulatory Compliance:* The regulatory landscape of medical devices and diagnostic tools is complex. Compliance with healthcare regulations and obtaining the necessary approvals may delay the deployment timeline and increase the cost of development.
4) *Ethical Issues:* AI in healthcare brings with it a number of ethical issues relating to patient privacy, security of data, and informed consent. The risk of breach of sensitive patient information and erosion of trust in the technology due to inadequate safeguards is also present.

VI. IMPLICATIONS AND CONSIDERATIONS FOR FUTURE WORK

The results of this work emphasize the transformative power of machine learning in lung cancer diagnosis, especially for early-stage diagnosis. The proposed framework using the advanced CNN architectures, like ResNet50, with feature extraction techniques such as DOST, offers a scalable and efficient alternative to traditional diagnostic methods. However, several implications and areas for improvement warrant further exploration: Clinical Implementation

Challenges One of the critical aspects is the integration of AIbased systems into clinical workflows. Although the model was able to achieve 98percent accuracy and 97percent F1-scores in controlled environments, its performance in clinical settings needs validation. Variability in imaging equipment, patient demographics, and healthcare infrastructure may impact the reliability of the system. Future research needs to focus on external validation by using diverse, real-world datasets.

Data Augmentation and Multimodal Integration While data augmentation techniques like rotation, flipping, and adding noise improve model performance, reliance on a single dataset, such as IQ-OTH/NCCD, limits scalability. Future studies should focus on collecting larger multimodal datasets with data from CT, PET, and histopathological images. This will not only improve generalizability of the models but also provide a more comprehensive understanding of tumor biology.

Explainability and Interpretability Despite achieving high diagnostic accuracy, deep learning models often behave as ”black boxes” and thus are not accepted widely by clinicians.

Recent work has highlighted the need for model interpretability, proposing techniques such as Grad-CAM and SHAP to visualize decision-making processes. Future work should include these techniques to build trust and allow clinicians to verify AI-driven predictions effectively.

Computational Efficiency The computational demands of deep learning models like ResNet50 pose challenges for deployment in resource-constrained settings. Researchers have explored lightweight architectures, such as MobileNet, for similar applications, though often at the expense of accuracy. Future research should aim to balance computational efficiency with predictive performance, ensuring accessibility across diverse healthcare systems.

Nanotechnology and Biomarkers Emerging technologies, including nanotechnology-based biomarkers such as miR-155 and Surface-Enhanced Raman Spectroscopy (SERS), may hold promise for non-invasive cancer detection. Combining these technologies with machine learning models could enhance early detection rates and reduce the number of invasive procedures. Future studies should be focused on hybrid approaches that combine imaging data with biomarker-based diagnostics. AI-driven diagnostics raise ethical concerns related to data privacy, algorithmic bias, and equitable access. Future research should address these issues by adhering to established ethical frameworks and incorporating fairness metrics into model evaluation. Collaboration with regulatory bodies will be essential to standardize AI implementations in healthcare. With these considerations in mind, future research can expand on the present framework to create stronger, more interpretable, and more accessible AI-driven diagnostic tools for lung cancer. This multidisciplinary approach will improve early detection, reduce mortality rates, and enhance patient outcomes globally.

VII. CONCLUSION

Here in this study, an overarching frame work is presented for machine-learning-based early-stage lung-cancer detection. Advanced CNN architectures such as ResNet50 and DOST, merged feature extraction with histogram equalization, yield a remarkable achievement from the proposed system in aspects like its accurate and reliable diagnosis. Coupling data augmentation with preprocessing addresses significant challenges of dataset imbalances and overfitting leading to making the model a better fit for real-life applications. The framework has a host of advantages over traditional diagnosis methods, including scalability and efficiency and potential real-time deployability. On the other hand, it also identifies crucial areas of future research, namely external validation on diverse datasets, model interpretability, and multimodal data sources.

Future advancements in nanotechnology, computational efficiency, and the ethical implementation of AI would be crucial to expanding the scope of machine learning in healthcare. This multidisciplinary approach may revolutionize early lung cancer detection, reduce mortality rates, and enhance patient outcomes across the world.

VIII. ACKNOWLEDGMENTS

We extend our gratitude to the faculty and staff of [Your Institution Name] for their support and guidance in the course of conducting this research. Special thanks are due to [Mentor/Advisor Name], whose invaluable guidance and constructive criticism helped form the study.

We also thank developers and maintainers of available public datasets, such as IQ-OTH/NCCD, that formed the core of this research. Great appreciation goes to them for adding to the scientific community's knowledge base. Lastly, we thank our peers and collaborators for the do the best in this work. This research is dedicated to improving healthcare solutions and making a positive impact on lung cancer detection and diagnosis.

REFERENCES

[1] Tasnim N., Noor K.R., Islam M., Huda M.N., Sarker I.H. A deep learning-based image processing technique for early lung cancer prediction. Proceedings of the 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), pp. 1060-1063, 2024.

[2] Kesav O.H., Rajini G.K. A systematic study on enhanced deep learning-based methodologies for detection and classification of early-stage cancers. Proceedings of the 2023 IEEE 5th International Conference on Cybernetics, Cognition, and Machine Learning Applications (ICCCMLA), pp. 328333, 2023.

[3] Abdullah M.F., Karim N.K.A., Sulaiman S.N., Shuaib I.L., Osman M.K., Alhamdu M.D.I. Classification of lung cancer stages from CT scan images using image processing and knearest neighbours. Proceedings of the 2020 IEEE Control and System Graduate Research Colloquium (ICSGRC), pp. 68-71, 2020.

[4] Romaszko A.M., Doboszynska A. Multiple primary´ lung cancer: A literature review. Advances in Clinical and Experimental Medicine, 27(5), pp. 725-730, 2018.

[5] Faisal M.I., Bashir S., Khan Z.S., Khan F.H. An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer. Proceedings of the 2018 International Conference on Emerging Trends in Engineering, Sciences, and Technology (ICEEST), pp. 1-4, 2018.

[6] Patra R. Prediction of lung cancer using machine learning classifiers. Computing Science, Communication, and Security: First International Conference (COMS2), pp. 132142, 2020.

[7] Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2097-2106, 2017.

[8] Chaturvedi P., Jhamb A., Vanani M., Nemade V. Prediction and classification of lung cancer using machine learning techniques. IOP Conference Series: Materials Science and Engineering, Vol. 1099, No. 1, p. 012059, 2021.

[9] Pawar V.J., Kharat K.D., Pardeshi S.R., Pathak P.D. Lung cancer detection system using image processing and machine learning techniques. Cancer Journal, 3(2020), p. 4, 2020.