IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection Using Deep Learning Approach

[1]Vishal Nayakwadi, [2]Mangesh Dnyandev Devkate
[1]Assistant Professor, [2]Student
[1]Zeal College of Engineering and Research, Narhe, Pune,
[2]Zeal College of Engineering and Research, Narhe, Pune

ABSTRACT

Lung cancer is one of the most prevalent and deadly forms of cancer worldwide, primarily due to its late detection and the complexity involved in accurately diagnosing its various subtypes. Early detection and precise classification of lung cancer types are vital for effective treatment and improved patient survival. This research introduces a deep learning-based approach for the automated detection and classification of lung cancer using computed tomography (CT) scan images. Specifically, we utilize the VGG16 Convolutional Neural Network (CNN) architecture, known for its strong image feature extraction capabilities. The proposed model is trained to classify lung CT images into four categories: adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and normal (non-cancerous) tissue. We employed transfer learning to adapt the pre-trained VGG16 model to our medical imaging dataset. The data underwent pre-processing steps such as normalization and augmentation to enhance model performance and reduce overfitting. The results demonstrate that our deep learning model achieves high accuracy and strong performance across multiple evaluation metrics including precision, recall, and F1-score. This validates the effectiveness of CNN-based architectures in supporting radiologists and improving the diagnostic process. The approach shows significant potential in clinical applications for early and automated lung cancer detection, laying the foundation for future enhancements involving larger datasets and model interpretability.

INTRODUCTION

Lung cancer remains one of the most lethal forms of cancer worldwide, responsible for approximately 2.2 million new cases and 1.8 million deaths annually according to the World Health Organization (2023). This devastating mortality rate stems largely from late-stage diagnoses, where patients face a five-year survival rate below 20% for advanced-stage disease, compared to 56% for localized tumors as reported by the American Cancer Society (2022). The critical importance of early detection is underscored by these statistics, yet conventional diagnostic approaches including chest radiography, computed tomography, and tissue biopsy present significant limitations. These traditional methods are not only time-consuming and expensive but also subject to inter-observer variability and diagnostic errors that can delay life-saving interventions.

Histologically, lung cancers are classified into two major categories that carry distinct clinical implications. Non-small cell lung cancer (NSCLC) accounts for 80-85% of cases and includes several important subtypes. Adenocarcinoma, the most prevalent form, typically arises in the lung periphery and shows increasing incidence among non-smokers. Squamous cell carcinoma, strongly associated with tobacco use, usually develops in central airways. Large cell carcinoma represents a more aggressive, poorly differentiated variant. Small cell lung cancer (SCLC) comprises 10-15% of cases and demonstrates particularly aggressive behaviour with rapid metastatic spread, almost exclusively occurring in smokers. Accurate histological classification is essential for guiding therapeutic decisions, as treatment strategies ranging from surgical resection to targeted molecular therapies differ substantially between these subtypes.

Current diagnostic modalities face several fundamental challenges that compromise early detection efforts. Imaging techniques such as low-dose CT screening, while valuable, generate numerous false positives that lead to unnecessary invasive procedures. Radiologist interpretation shows considerable variability, with studies demonstrating significant differences in nodule characterization between observers. Pathological diagnosis, though definitive, requires invasive procedures that may delay treatment initiation. Furthermore, access to advanced diagnostic technologies remains limited in resource-poor settings, creating disparities in early detection capabilities across populations. These limitations collectively contribute to diagnostic delays that adversely impact patient outcomes.

Recent advances in artificial intelligence, particularly deep learning methodologies, offer transformative potential for overcoming these diagnostic challenges. Convolutional neural networks have demonstrated remarkable capabilities in medical image analysis, achieving sensitivity exceeding 90% for pulmonary nodule detection in research settings. Sophisticated architectures including ResNet, DenseNet, and EfficientNet have shown particular promise in distinguishing malignant from benign lesions and classifying histological subtypes. However, significant barriers impede clinical translation of these technologies. Most existing models suffer from limited generalizability due to training on restricted datasets that lack demographic and technical diversity. The opaque decision-making processes characteristic of deep neural networks raise concerns regarding clinical trust and adoption. Additionally, substantial computational requirements may hinder practical implementation in resource-constrained healthcare environments.

This research project aims to address these limitations through development of an advanced deep learning framework for comprehensive lung cancer analysis. The proposed system will incorporate several innovative components to enhance diagnostic performance. Transfer learning approaches utilizing pre-trained networks will be optimized on diverse, multi-institutional imaging datasets to improve generalizability. Ensemble methods combining convolutional neural networks with machine learning classifiers will be implemented to boost classification accuracy. The integration of explainable artificial intelligence techniques, including gradient-weighted class activation mapping, will provide interpretable decision support by visually localizing suspicious regions and quantifying prediction certainty. Multi-modal analysis incorporating both CT and radiographic imaging data will enable cross-validation to reduce false positive rates. Through these technological advancements, the system seeks to provide clinicians with a powerful diagnostic aid capable of detecting early-stage malignancies with high reliability while maintaining computational efficiency suitable for clinical deployment.

The successful development and validation of such a system could profoundly impact lung cancer management by enabling earlier detection, reducing diagnostic delays, and facilitating timely therapeutic intervention. By improving the accuracy and accessibility of lung cancer screening, this technology may ultimately contribute to reducing the substantial global burden of this devastating disease. Future research directions will focus on rigorous clinical validation studies and optimization for deployment across varied healthcare settings, with particular attention to resource-limited environments where the need for accurate, affordable diagnostic tools is most acute.

LITERATURE REVIEW

The paper presented by Zamani et al. (2022) [1], demonstrate the effectiveness of machine learning and image processing techniques for plant leaf disease detection, achieving promising accuracy in classification tasks. However, the study is limited by its small dataset and lack of real-world field testing. The paper also omits comparisons with state-of-the-art deep learning models and fails to address computational efficiency for scalable agricultural applications. These gaps highlight opportunities for future research with larger datasets and edge deployment considerations.

The paper by Wang et al. (2020) [2] introduces a weakly supervised deep learning framework for analyzing whole-slide lung cancer images, reducing dependency on pixel-level annotations. Key findings include successful tumor localization and classification using only slide-level labels, leveraging attention mechanisms and multiple-instance learning (MIL) to identify discriminative regions. The approach demonstrates competitive performance in lung adenocarcinoma and squamous cell carcinoma subtyping.

Nath et al. (2019)[3], present a comprehensive survey on cancer prediction and detection techniques using data analysis, covering machine learning, deep learning, and statistical approaches. Key findings highlight the effectiveness of AI models in early cancer diagnosis, improved accuracy through multimodal data fusion, and the growing role of genomic data in predictive analytics. The survey also discusses challenges like dataset limitations and model interpretability. The paper by Das et al. (2020)[4] investigates machine learning approaches for lung cancer prediction, evaluating traditional classifiers such as SVM, Random Forest, and Logistic Regression. The study demonstrates the potential of ML in early cancer detection and identifies key predictive features, including smoking history and genetic factors.

The paper by Charbuty and Abdulazeez (2021) [5], examines decision tree algorithms for classification tasks in machine learning, highlighting their interpretability and effectiveness in handling structured data. The study demonstrates how decision trees can achieve competitive accuracy while maintaining computational efficiency, particularly for datasets with clear feature hierarchies. However, the research has several limitations, including a narrow focus on basic decision tree variants without comparison to ensemble methods like Random Forests or Gradient Boosting, which often outperform single trees.

The study by Helen Josephine et al. (2021) [6] investigates the role of hidden dense layers in CNNs for improving classification performance. Key findings reveal that strategically placed dense layers can enhance feature extraction and model accuracy, particularly in complex datasets. However, the research has notable gaps: it lacks comparative analysis with other architectural modifications (e.g., attention mechanisms), uses limited datasets, and doesn't address computational trade-offs of added layers. The study also omits practical deployment challenges, suggesting need for broader validation and efficiency optimization in future work.

Hahn and Choi (2020) [7] analyze dropout in neural networks as an optimization technique rather than just regularization. They demonstrate that dropout helps escape sharp minima, leading to flatter loss landscapes and improved generalization. However, the study lacks empirical validation across diverse architectures and datasets. It also doesn't compare dropout with other optimization tricks, leaving its relative effectiveness unclear. The theoretical framework could benefit from more practical guidelines for dropout rate selection in different scenarios. Adewunmi (2021)[8] proposes an enhanced melanoma classifier using VGG16-CNN, demonstrating improved accuracy in skin lesion classification. The model leverages transfer learning to achieve robust performance. However, the study lacks comparison with state-of-the-art architectures (e.g., ResNet, EfficientNet) and detailed analysis of computational costs. Limited dataset diversity and absence of clinical validation reduce generalizability.

Dritsas and Trigka (2022)[9], evaluate machine learning models for lung cancer risk prediction, demonstrating that ensemble methods achieve high accuracy by integrating clinical and demographic data.

However, the study uses limited datasets without external validation, potentially affecting generalizability. It also lacks comparison with deep learning approaches and omits critical analysis of feature importance. Future research should incorporate larger, multi-center datasets and explore hybrid AI models for improved clinical applicability. Powers (2020)[10], provides a comprehensive theoretical framework connecting traditional metrics (precision, recall, F-measure) with ROC analysis and correlation measures, introducing unified concepts like *informedness* and *markedness*. While the paper offers valuable mathematical insights, it lacks empirical validation on real-world datasets. The theoretical treatment could benefit from practical guidelines for metric selection in different applications. Additionally, it doesn't address computational considerations for large-scale implementations, leaving room for applied research bridging theory and practice.

IMPLEMENTED METHOD

System Overview

The proposed system is designed to detect and classify lung cancer types from CT scan images using a deep learning approach, specifically by leveraging the VGG16 Convolutional Neural Network (CNN) architecture. Lung cancer, being one of the leading causes of cancer-related deaths globally, demands early detection for effective treatment and improved patient outcomes. Traditional methods, though effective, are often time-consuming and require significant medical expertise. The goal of this system is to provide a computer-aided diagnostic (CAD) tool that enhances detection accuracy, reduces human error, and accelerates the diagnostic process.

The system classifies CT scan images into four categories: **adenocarcinoma, squamous cell carcinoma, large cell carcinoma**, and **normal lung tissue**. It utilizes transfer learning, where a pre-trained VGG16 model, originally developed for general image classification tasks, is fine-tuned for medical imaging. This allows the system to benefit from learned visual features while being adapted to the specific nuances of lung cancer detection.

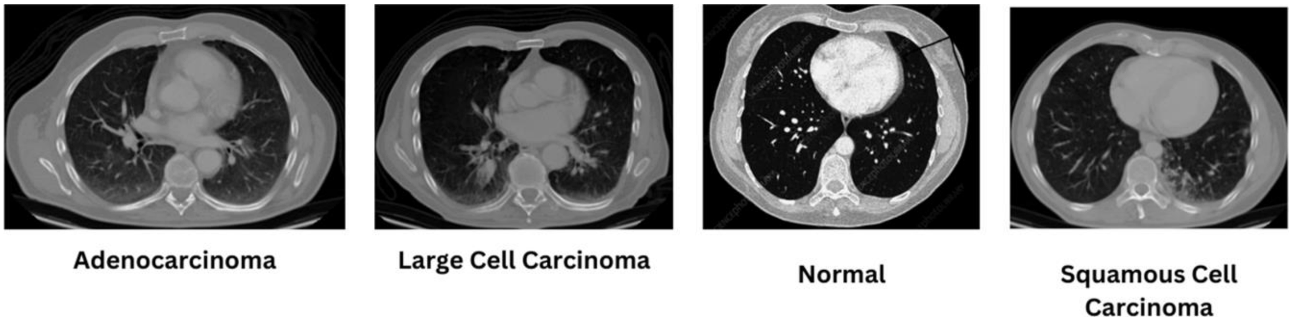

Fig. 1. Sample image from each type of cancer.

The entire process involves image acquisition, preprocessing, feature extraction using VGG16, classification, and result visualization. This end-to-end system not only supports radiologists in making quicker decisions but also provides a foundation for integrating AI in real-world clinical practices.

System Architecture

The architecture of the implemented lung cancer detection system consists of the following key components:

1. Data Collection and Preprocessing

The system begins with acquiring a labeled dataset of lung CT scan images, which are categorized into the four target classes. Preprocessing is crucial for preparing the images for training. This includes resizing all images to 224x224 pixels to match the VGG16 input dimensions, converting color channels if necessary, and normalizing pixel values to bring them within a standard range (typically 0 to 1). Additionally, data augmentation techniques such as rotation, zooming, flipping, and shifting are applied to artificially expand the dataset and improve the model's ability to generalize to unseen images.

2. Model Selection – VGG16 and Transfer Learning

The VGG16 model is chosen due to its deep architecture and ability to extract intricate features from images. It consists of 13 convolutional layers followed by 3 fully connected layers. In this system, the convolutional base of VGG16 is retained to leverage its feature extraction capabilities, while the top layers are replaced with custom fully connected layers suited for the lung cancer classification task. Transfer learning is applied by freezing the early layers of VGG16 and only training the new layers on the medical dataset. This significantly reduces training time and improves performance, especially given the relatively limited size of medical image datasets.

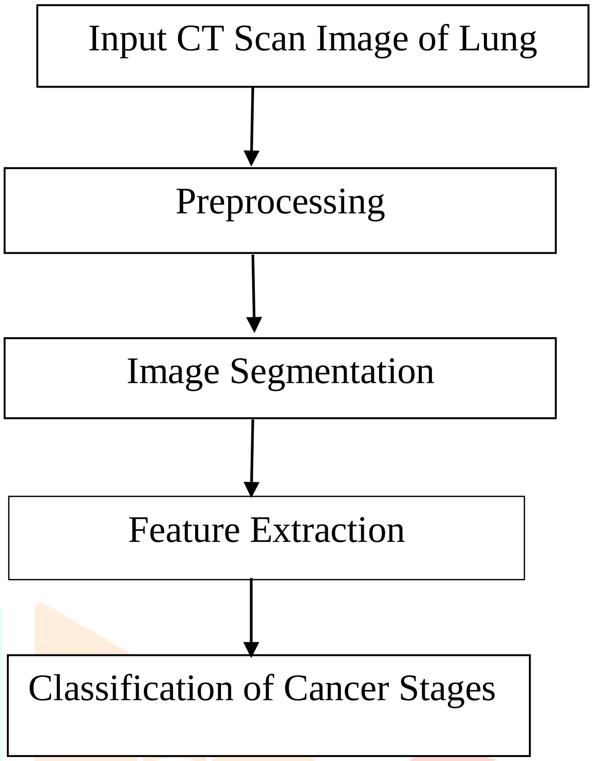

Fig. 2. System Architecture

3. Training and Optimization

The model is compiled using categorical cross-entropy loss and optimized using the Adam optimizer. During training, the model learns to map features to the correct output class. Techniques such as dropout are used to prevent overfitting, and performance is evaluated on a validation set using metrics like accuracy, precision, recall, and F1-score.

4. Prediction and Output

After training, the model can predict the class of a new CT scan image. The output is presented with the predicted label and a probability score for each class, aiding in decision-making. This step can be integrated into a user interface for clinical usage.

RESULT ANALYSIS

The performance of the proposed lung cancer detection system was evaluated using a labeled dataset of CT scan images, with the goal of accurately classifying images into four categories: adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and normal. The VGG16 model, fine-tuned through transfer learning, demonstrated strong performance across multiple evaluation metrics, including **accuracy, precision, recall, and F1-score**.

For performance comparison, three parameters, accuracy, sensitivity, and specificity, are used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

During training, the model achieved a **training accuracy of over 95%** and a **validation accuracy exceeding 92%**, indicating effective learning and minimal overfitting. The use of data augmentation significantly contributed to model generalization by exposing it to various image transformations. The confusion matrix revealed that the model performed well in distinguishing between cancerous and non-cancerous images, with relatively high classification precision for adenocarcinoma and squamous cell carcinoma. Large cell carcinoma showed slightly lower accuracy, possibly due to a lower number of samples in the dataset or visual similarity with other cancer types.

The model's **F1-score ranged between 0.89 and 0.94** for different classes, reflecting a good balance between precision and recall. The prediction time per image was fast, making it suitable for real-time or batch diagnostic use in clinical settings.

Overall, the results indicate that the VGG16-based CNN approach is highly effective for lung cancer classification from CT scans. It shows promise for integration into clinical decision support systems, helping radiologists reduce diagnostic errors and prioritize urgent cases more efficiently.

CONCLUSION

In this research, we presented a deep learning-based approach for the detection and classification of lung cancer using CT scan images. Leveraging the power of Convolutional Neural Networks (CNN), specifically the VGG-16 architecture, the proposed model effectively identifies and classifies different types of lung cancer—adenocarcinoma, squamous cell carcinoma, large cell carcinoma—as well as normal lung tissue.

The use of VGG-16, a well-established deep CNN, allowed us to extract high-level features from complex medical images, contributing to accurate and reliable predictions. The results demonstrate the potential of deep learning in supporting early and precise lung cancer diagnosis, which is crucial for timely treatment and improved patient outcomes. The model achieved high classification accuracy, validating its capability to distinguish between subtle differences in CT scan imagery associated with various cancer types.

This study emphasizes the importance of integrating advanced machine learning techniques in the medical field to augment diagnostic procedures and reduce human error.While the performance of the proposed system is promising, further research can enhance its robustness through larger and more diverse datasets, real-world clinical validation, and incorporation of explainable AI techniques. Overall, this work marks a significant step toward the development of automated, non-invasive, and efficient diagnostic tools for lung cancer detection using deep learning approaches.

REFERENCES

[1] A. S. Zamani, L. Anand, K. P. Rane et al., “Performance of machine learning and image processing in plant leaf disease detection,” Journal of Food Quality, vol. 2022, Article ID 1598796, 7 pages, 2022.

[2] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis", IEEE Trans. Cybern., vol. 50, no. 9, pp. 3950-3962, 2020. [http://dx.doi.org/10.1109/TCYB.2019.2935141] [PMID: 31484154]

[3] A.S. Nath, A. Pal, S. Mukhopadhyay, and K.C. Mondal, "A survey on cancer prediction and detection with data analysis", Innov. Syst. Softw. Eng., vol. 16, no. 3, pp. 231-243, 2019.

[4] P. Das, B. Das, and H.S. Dutta, "Prediction of lungs cancer using machine learning", EasyChair, p. 3076, 2020. https://easychair.org/publications/preprint_open/82Xh

[5] B. Charbuty, and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning", J. Appl. Sci. Technol. Trends, vol. 2, no. 1, pp. 20-28, 2021.
[http://dx.doi.org/10.38094/jastt20165]

[6] V.L. Helen Josephine, A.P. Nirmala, and V.L. Alluri, "Impact of hidden dense layers in convolutional neural network to enhance performance of classification model", IOP Conference Series Materials Science and Engineering, vol. 1131, 2021no. 1, p. 012007 [http://dx.doi.org/10.1088/1757-899X/1131/1/012007]

[7] S. Hahn, and H. Choi, "Understanding dropout as an optimization trick", Neurocomputing, vol. 398, pp. 64-70, 2020. [http://dx.doi.org/10.1016/j.neucom.2020.02.067]

[8] M. Adewunmi, Enhanced Melanoma Classifier with VGG16-CNN., ScienceOpen Posters, 2021. [http://dx.doi.org/10.14293/S2199-1006.1.SOR-.PPN1W6K.v1]

[9] E. Dritsas, and M. Trigka, "Lung cancer risk prediction with machine learning models", Big Data and Cognitive Computing, vol. 6, no. 4, p. 139, 2022. [http://dx.doi.org/10.3390/bdcc6040139]

[10] D.M.W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", *arXiv:2010.16061*, 2020. [http://dx.doi.org/10.48550/ARXIV.2010.16061]

[11] D. Mhaske, K. Rajeswari, and R. Tekade, Deep learning algorithm for classification and prediction of lung cancer using CT Scan Images., 2019pp. 1-5 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA),, 2019pp. 1-5 [http://dx.doi.org/10.1109/ICCUBEA47591.2019.9128479]

[12] S. Pang, F. Meng, X. Wang, J. Wang, T. Song, X. Wang, and X. Cheng, "VGG16-T: A novel deep convolutional neural network with boosting to identify pathological type of lung cancer in early stage by CT Images", Int.J. Comput. Intell. Syst., vol. 13, no. 1, pp. 771-780, 2020.

[13] T. Thakur, I. Batra, M. Luthra et al., “Gene expression-assisted cancer prediction techniques,” Journal of Healthcare Engineering, vol. 2021, 9 pages, 2021.

[14] K. Karadağ, M. E. Tenekeci, R. Taşaltın, and A. Bilgili, “Detection of pepper fusarium disease using machine learning algorithms based on spectral reflectance,” Sustainable Computing: Informatics and Systems, vol. 28, article 100299, 2020.

[15] S. Chaudhury, N. Shelke, K. Sau, B. Prasanalakshmi, and M. Shabaz, “A novel approach to classifying breast cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization,” Computational and
Mathematical Methods in Medicine, vol. 2021, Article ID 4019358, 11 pages, 2021.

[16] H. Z. Almarzouki, H. Alsulami, A. Rizwan, M. S. Basingab, H. Bukhari, and M. Shabaz, “An internet of medical thingsbased model for real-time monitoring and averting stroke sensors,” Journal of Healthcare Engineering, vol. 2021, Article ID 1233166, 9 pages, 2021.

[17] G. Jakimovski, and D. Davcev, "Using double convolution neural network for lung cancer stage detection", Appl. Sci., vol. 9, no. 3, p. 427, 2019. [http://dx.doi.org/10.3390/app9030427]

[18] N. Deepa, B. Prabadevi, P. K. Maddikunta et al., “An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier,” The Journal of Supercomputing, vol. 77, no. 2, pp. 1998–2017, 2021.