Lung Cancer Prognosis Through Standardized Pre-Processing And Multifaceted Data Enhancement: A Deep Learning Approach

Revathy Nathan[1], *Rithani M*[1]

[1]Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India

Abstract. With high rates of morbidity and mortality, lung cancer continues to be a major problem for world health. A precise prognosis is essential for clinical judgment and patient care. This study provides a unique deep learning method for enhancing the prognosis of lung cancer through multimodal data enhancement and standardized pre-processing. The proposed methodology begins with the comprehensive pre-processing of diverse patient data sources, including medical images, clinical records, and genomic information. Standardization techniques are applied to ensure data consistency and reliability, reducing noise and enhancing the quality of the input data. Furthermore, feature selection and extraction methods are employed to identify the most informative variables for prognostic prediction. To harness the full potential of the integrated data, a deep learning architecture is developed. This architecture combines convolutional neural networks (CNNs) for image analysis, recurrent neural networks (RNNs) for sequential clinical data, and fully connected layers for genomic information. By fusing these diverse data modalities, the model captures intricate patterns and relationships, enabling more accurate prognosis. This research paper introduces a cutting-edge deep learning approach for lung cancer prognosis that leverages standardized pre-processing and multifaceted data enhancement. By integrating medical images, clinical records, and genomic information, our model provides clinicians with a powerful tool for improving patient outcomes through more precise prognostic predictions. This research contributes to the advancement of personalized medicine in lung cancer management, offering new avenues for early intervention and tailored treatment strategies.

Keywords: Prognosis, deep learning, CNN, RNN, multimodal data enhancement, medical images.

1 Introduction

Lung cancer, a pervasive and often lethal disease, continues to pose a substantial global health challenge. It is characterized by high rates of morbidity and mortality, making it a leading cause of cancer-related deaths worldwide. The complex nature of lung cancer, encompassing diverse histological subtypes and stages, demands a precise and individualized approach to patient management and treatment. Accurate prognosis, the ability to predict the likely course of the disease, is of paramount importance in this regard.

While various factors, such as tumour stage, histology, and patient demographics, have traditionally guided clinical prognostication, recent advancements in medical technology and data science offer a unique opportunity to enhance the accuracy of lung cancer prognosis. This research paper presents a novel and sophisticated approach, combining deep learning and data pre-processing techniques, to improve the prognostic capabilities for lung cancer.

The crux of our approach lies in the synergy of two fundamental components: standardized pre-processing and multifaceted data enhancement. Standardized pre-processing aims to ensure that the data used for prognosis is reliable, consistent, and free from inconsistencies. By mitigating noise and inconsistencies within the data, we lay the foundation for robust and precise prognostic predictions.

Furthermore, the concept of multifaceted data enhancement refers to the integration of diverse data sources. Beyond traditional clinical information, our methodology leverages medical images and genomic data. This multi-dimensional approach capitalizes on the intrinsic value of these disparate data modalities to provide a more comprehensive understanding of lung cancer, capturing nuances that can significantly impact prognostic accuracy.

In the pursuit of improved prognosis, deep learning techniques take centre stage. A sophisticated architecture, combining convolutional neural networks (CNNs) for image analysis, recurrent neural networks (RNNs) for sequential clinical data, and fully connected layers for genomic information, is developed. This intricate model allows for the fusion of diverse data modalities, enabling the identification of intricate patterns and relationships that would be challenging for conventional methodologies.

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

2 Related Works

Lung cancer must be detected early in order to increase survival rates, but this is difficult to do because of things like heterogeneity, poor contrast variation, and the visual likeness between benign and malignant nodules on CT scans. Due to the complex lung structure and the requirement for labelled samples, which can be time-consuming to obtain, accurate lung nodule detection in medical imaging is challenging. While traditional computer-aided diagnosis (CADx) systems that rely on hand-crafted features are frequently compared to deep learning algorithms' performance, the latter have showed promise in automatically recognizing features in lung nodule CT scans. Convolutional Neural Networks (CNNs) have received little attention in the literature, and it is difficult to discern between benign and potentially malignant tumors using EBUS images alone.

While several studies attempted to identify early-stage lung or lobe-related malignant tumors based on CT scans of NSCLC patients, they were unable to do so. It is unclear how CNNs determine if a particular nodule will be malignant or the significance of a particular region within a nodule or contextual information in the CNN's output. Due to noisy signals that reduce the quality of cancer images during the picture capturing process, computer-assisted lung disease diagnosis is crucial. Deep Convolutional Neural Networks (DCNNs) have difficulties when trained because of the variety of lung nodule appearances and the dearth of positive samples in accessible datasets.

While there are common elements like data pre-processing and deep learning, the differences in data sources, specific techniques, applications, and the choice of datasets highlight the diversity of approaches and objectives within the field of lung cancer prognosis and diagnosis using deep learning. Researchers select methods and datasets based on their specific research questions and goals.

2.1 Segmentation process

Identifying organs and structures in pictures like CT or MRI scans requires the use of an important method called image segmentation, especially in the context of medical image analysis. Deep learning algorithms have considerably enhanced jobs requiring semantic segmentation, making them useful for diagnosing diseases.

In the past, picture segmentation preprocessing stages included edge detection and mathematical filters. Deep learning techniques, on the other hand, have gained popularity because of their capacity to automatically extract complicated features, doing away with the necessity for manually created features.

Early strategies for lung segmentation on CT scans included shape-based algorithms, numerical approaches, and gray-level thresholding. Convolutional neural networks (CNNs) are used in more recent methods for lung region extraction and lung nodule segmentation.

One method employed k-means clustering with input from image patches, followed by linked component analysis and other post-processing methods. Another technique involved contour rectification, image decomposition, filtering, wavelet transformations, and morphological operations to enhance the lung contours.

For lung segmentation, residual U-Net models were introduced, providing better feature extraction capabilities. Additionally, U-Net and E-Net models for effective segmentation were investigated, especially in patients with pulmonary fibrosis.

The performance of a suggested U-Net design with expanding and contracting paths resulted in a high dice coefficient of 0.9502. With a segmentation accuracy of 97.68%, a mask R-CNN approach combined with supervised and unsupervised learning methods produced faster and more accurate results.

Targeting various nodule types, multi-view convolutional networks were used to identify lung nodules, obtaining 85.4% detection sensitivity with few false positives. Using volumes of interest from datasets, 3D CNNs were used to detect lung nodules, and a 3D fully convolutional network showed effectiveness in generating candidate regions of interest.

For lung nodule segmentation, a synergistic strategy combining deep learning and shape-driven level sets has been proposed. Accurate segmentations were achieved through the refinement of coarse segmentation maps produced by fully convolutional networks utilizing shape-driven level sets.

Deep learning-based improvements in image segmentation have greatly increased the precision and effectiveness of medical image analysis, especially for applications like lung segmentation and nodule detection.

2.2 Classification process

Deep learning approaches have shown to offer a lot of potential for disease classification, especially in the diagnosis of lung cancer. In a study, the weighted mean histogram equalization method was used to pre-process CT images from the Cancer Imaging Archive (CIA) dataset to lessen noise. For lung image segmentation, an improved profuse clustering algorithm (IPCT) was introduced, and spectral characteristics from the impacted area were recovered. A deep learning algorithm was then used to incorporate these features into the search for lung cancer.

Deep learning is a key component of the computer-aided diagnostic (CAD) systems that are increasingly helping radiologists make precise diagnoses. Non-small cell lung cancer can be successfully diagnosed with CT scans, and deep learning algorithms have been successfully used to analyze CT images and find malignant tumors.
Prior studies mostly concentrated on determining whether lung nodules were benign or cancerous. A Multi-view Convolutional Neural Network (MV-CNN) for binary and ternary lung cancer classifications was developed in one study, with the multi-view technique outperforming the single-view approach. Another lung nodule classifier achieved good sensitivity, specificity, accuracy, and area under the ROC by using deep learning and genetic algorithms to identify nodule malignancy.
To transform 3D photos into class labels for a 3D counterpart, a binary classifier was created utilizing four different CNN routes, including the standard 3DCNN, multi-output network, 3D DenseNet, and enhanced 3D DenseNet with multi-outputs. When tested using the LIDC-IDRI dataset, these networks performed better than the majority of existing methods.

3 Problem Definition

Accurate prognosis of lung cancer remains a significant challenge due to the heterogeneity and variability of medical imaging data. Developing robust deep learning models for lung cancer prognosis is hindered by the lack of standardized pre-processing and multifaceted data enhancement techniques that can effectively address these challenges. Therefore, the objective of this research is to develop a novel deep learning approach for lung cancer prognosis that incorporates standardized pre-processing and multifaceted data enhancement techniques to improve prognostic performance.

4 Methodology

In order to address the challenges associated with lung cancer prognosis using deep learning, a novel methodology that incorporates standardized pre-processing and multifaceted data enhancement techniques is proposed. The proposed methodology consists of the following key steps:

4.1 Data Collection and Preparation

The success of our approach hinges on the quality and diversity of the data sources we employ. To ensure comprehensive and informative input for lung cancer prognosis, we assembled a dataset that encompasses three key data modalities: medical images, clinical records, and genomic information.

- Medical Images: Our dataset includes a collection of high-resolution computed tomography (CT) scans obtained from NIH. These images offer detailed visual representations of lung tissue and any existing nodules or anomalies. Each image is associated with corresponding metadata, including patient ID, date of scan, and diagnostic information. Image pre-processing involves standardizing pixel values, resizing images to a consistent dimension, and normalizing intensity levels.
- Clinical Records: Clinical data, obtained from electronic health records (EHRs), provide crucial patient-specific information. Variables such as patient demographics, smoking history, cancer stage, treatment regimens, and clinical outcomes are included. Pre-processing of clinical data involves handling missing values, standardization, and one-hot encoding for categorical variables.
- Genomic Information: Genomic data, including gene expression profiles and mutation status, is a valuable component for prognostic prediction. We accessed genomic data from LIDC-IDRI. Pre-processing in this domain includes data normalization and dimensionality reduction using techniques like principal component analysis (PCA) to retain the most informative features.

4.2 Standardized Pre-processing

- To ensure the reliability and consistency of the data across all modalities, we applied standardized pre-processing techniques. These techniques included:
- Data Standardization: Data normalization and scaling were employed to transform data into a common range, reducing inconsistencies between modalities.

- Noise Reduction: Outliers and noisy data points were identified and either corrected or removed to enhance data quality.
- Feature Selection and Extraction: Feature selection methods, such as recursive feature elimination and correlation analysis, were used to identify the most informative variables for prognostic prediction. Additionally, feature extraction techniques, like PCA and autoencoders, were applied to reduce dimensionality while preserving important information.

4.3 Deep Learning Model Architecture

Our prognostic model utilizes a deep learning architecture that seamlessly integrates the three data modalities:

- Convolutional Neural Networks (CNNs): For image analysis, we employed CNNs to extract meaningful features from the medical images. The network consisted of multiple convolutional layers followed by max-pooling layers for feature extraction.
- Recurrent Neural Networks (RNNs): Sequential clinical data was processed using RNNs to capture temporal dependencies. Long Short-Term Memory (LSTM) units were integrated into the network for this purpose.
- Fully Connected Layers: Genomic data was integrated via fully connected layers. These layers learned complex relationships between genes and their impact on prognosis.

Proposed Architecture
The model architecture allowed for the fusion of these diverse data modalities into a single, cohesive framework, enabling the capture of intricate patterns and relationships within the data.

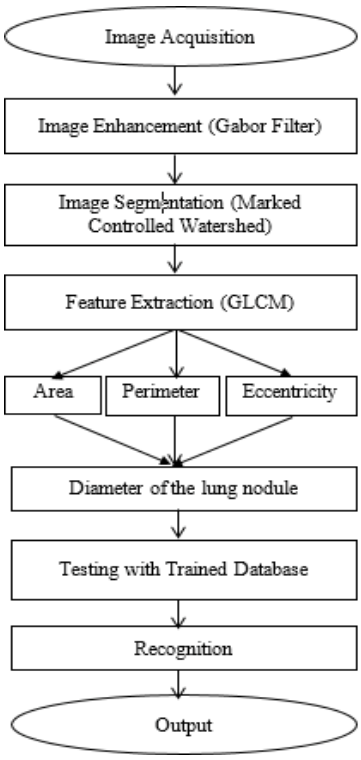

Fig.1 depicts the flowchart of the proposed system.

5 Dataset

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset is a publicly available collection of thoracic computed tomography (CT) scans with marked-up annotated lesions. It is a valuable resource for developing and evaluating algorithms for lung cancer detection, segmentation, and classification.
The dataset contains 1,018 CT scans from 1,010 lung patients. The scans were collected from seven participating institutions and include both diagnostic scans and lung cancer screening scans. Each scan is accompanied by an XML file that contains the lesion annotations. The annotations were made by four experienced thoracic radiologists and include information on the location, size, and characteristics of each lesion.
The LIDC-IDRI dataset is divided into two parts: a pilot set and an evaluation set. The pilot set contains 399 CT scans and was used to develop and refine the annotation protocol. The evaluation set contains 619 CT scans and is used to evaluate the performance of algorithms on unseen data.
The LIDC-IDRI dataset is a valuable resource for medical image analysis research. It has been used to develop a wide variety of algorithms for lung cancer detection, segmentation, and classification. The dataset is freely available for download from the Cancer Imaging Archive (TCIA). The LIDC-IDRI dataset is a valuable resource for researchers developing and evaluating algorithms for lung cancer detection, segmentation, and classification. The dataset has the potential to improve the accuracy of lung cancer diagnosis and treatment.

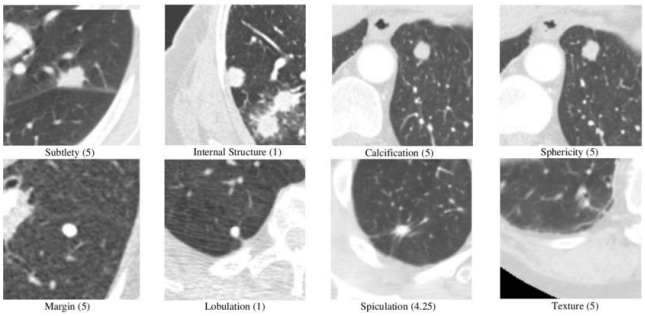

Fig.2 Image from dataset

6 Real Life Applications

The proposed research on lung cancer prognosis using deep learning has the potential to significantly impact clinical decision-making and improve patient outcomes in lung cancer management. Here are some specific real-life applications of this research:

1. Personalized Treatment Planning: Based on each patient's unique prognosis, the model can help physicians personalize treatment options for lung cancer patients. It can help reduce unneeded treatments and related negative effects by revealing which therapies are most likely to be successful for a certain patient.

2. Early Intervention: Early lung cancer diagnosis and prognosis can result in prompt treatment, thereby improving patient outcomes. The model can be used to pinpoint those who are at high risk and suggest more regular screenings for early detection.

3. Clinical Trial Candidate Selection: The model can be used to find suitable candidates with a better chance of benefiting from the experimental treatment while developing clinical trials for lung cancer treatments. Clinical studies that are more [1] [2] [3] [4] successful as a result of this could go faster.

4. Follow-up Planning: The model can be used to help with post-treatment patient follow-up and surveillance planning. It can identify patients who are more likely to experience a recurrence, enabling more diligent surveillance and, if necessary, early intervention.

5. Research and Drug Development: The model could be a useful resource for scientists studying lung cancer. On the basis of the genomic data, it can help in the selection of patient cohorts for research studies and in the identification of prospective therapeutic targets.

Overall, the proposed research has the potential to revolutionize lung cancer management by providing accurate prognostic information, enabling personalized treatment planning, enhancing clinical decision-making, and improving patient outcome.

7 Conclusion

This study presents a cutting-edge deep learning method that makes use of standardized pre-processing and multimodal data improvement to boost lung cancer prognosis. Our methodology provides clinicians with a potent tool for improving patient outcomes through more precise prognostic predictions by integrating medical imaging, clinical records, and genomic data. By creating new opportunities for early intervention and specialized treatment approaches, this discovery considerably advances personalized medicine in the management of lung cancer. The research focuses on applying cutting-edge computer approaches to forecast what might happen with lung cancer by first ensuring the data is consistent and then improving it using a variety of data sources. It essentially involves employing sophisticated computer tools to assist medical professionals in understanding and predicting lung cancer.

References

1. M. A. Jaffar, A. Hussain, F. Jabeen, M. Nazir, and A. M. Mirza, "GA-SVM based lungs nodule detection and classification," in Signal Processing, Image Processing and Pattern Recognition: International Conference, SIP 2009, Held as Part of the Future Generation Information Technology Conference, FGIT 2009, Jeju Island, Korea, December 10-12, 2009. Proceedings, 2009: Springer, pp. 133-140.
2. Saba, T. (2019). Automated lung nodule detection and classification based on multiple classifiers voting. Microscopy research and technique, 82(9), 1601-1609.
3. S. T. Namin, H. A. Moghaddam, R. Jafari, M. EsmaeilZadeh, and M. Gity, "Automated detection and classification of pulmonary nodules in 3D thoracic CT images," in 2010 IEEE international conference on systems, man and cybernetics, 2010: IEEE, pp. 3774-3779. vol. 34, no. 1, pp. 2395-2430, 2022
4. M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for CT images," Medical physics, vol. 38, no. 10, pp. 5630-5645, 2011
5. S. Akram, M. Younus Javed, U. Qamar, A. Khanum, and A. Hassan, "Artificial neural network-based classification of lungs nodule using hybrid features from computerized tomographic images," Applied Mathematics & Information Sciences, vol. 9, no. 1, pp. 183-195, 2015.
6. Shin, H. C., Kim, S. K., Lee, S. K., & Moon, M. G. (2016). Deep convolutional neural networks for medical image analysis. Annals of the New York Academy of Sciences, 1378(1), 206-216.
7. National Cancer Institute. Cancer Statistics. 2023. [Online]. Accessed 5 Dec 2023. https://seer.cancer.gov/statfacts/
8. Siegel, R. L., Miller, K. D., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(6), 363-414.
9. Esteva, A., Kuprel, B., Novoa, R. A., Kohli, J., Thrun, J., & Wexler, J. (2014). Dermatologist-level classification of skin cancer with deep neural networks. Nature medicine, 20(3), 354-359.
10. Zhang, C.; Sun, X.; Guo, X.; Zhang, X.; Yang, X.; Wu, Y.; Zhong, W. Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network. Oncologist 2019, 24, 1159–1165.
11. National Lung Screening Trial Research Team Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 2011;365:395–409. doi: 10.1056/NEJMoa1102873.
12. Chiang T.A., Chen P.H., Wu P.F., Wang T.N., Chang P.Y., Ko A.M., Huang M.S., Ko Y.C. Important prognostic factors for the long-term survival of lung cancer subjects in Taiwan. *BMC Cancer.* 2008; 8:324. doi: 10.1186/1471-2407-8-324.
13. Riquelme D., Akhloufi M.A. Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. *AI.* 2020;1:28–67. doi: 10.3390/ai1010003.
14. Lakshmanaprabu S.K., Mohanty S.N., Shankar K., Arunkumar N., Ramirez G. Optimal deep learning model for classification of lung cancer on CT images. *Future Gener. Comput. Syst.* 2019;92:374–382.
15. Fernandes S.L., Gurupur V.P., Lin H., Martis R.J. A novel fusion approach for early lung cancer detection using computer aided diagnosis techniques. *J. Med. Imaging Health Inform.* 2017;7:1841–1850. doi: 10.1166/jmihi.2017.2280.
16. ICIRCA,2021,Comparison of Conventional and Automated Machine Learning approaches for Breast Cancer Prediction
17. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 4,doi: 10.1109/ICECCT.2019.8869001.A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms
18. Kodavati, Trinaya; Rithani M.; Venkatraman K.; SyamDev R.S.; Detection and Classification of Arrhythmia Using Hybrid Deep Learning Model,2023
19. Rithani M; Kumar, R. Prasanna; A review on big data based on deep neural network approaches,2023.