Machine Learning Approaches to Lung Cancer Prediction: A Comparative Study

Sunita Ranadhir Landge[1*] and Dinesh Jain[2]

[1]Ph.D. Scholar, SAGE University, Indore, India
[2]Professor, SAGE University, Indore, India

*Corresponding author e-mail:sunitarlandge@gmail.com

Abstract. Lung cancer continues to rank among the most common and fatal types of cancer worldwide, significantly affecting both patient quality of life and public health. Improving treatment results and survival rates requires early detection and precise diagnosis. In recent years, deep learning algorithms have shown promise in increasing the accuracy of lung cancer prediction. Large datasets are needed for deep learning (DL) in order to train the model. However, the sample size of the current dataset is small, which limits the model's generalizability. We conduct a comparative analysis for lung cancer forecasts in this research. The model was trained using DL algorithms such as recurrent neural networks (RNN), convolutional neural networks (CNN), and artificial neural networks (ANN), SVM (Support vector machine) and VGG-16.

Keywords:
Lungs cancer, deep learning, convolutional neural network, artificial neural network, **recurrent neural networks**

1. Introduction

Lung diseases represent a significant global health concern, contributing substantially to illness and mortality rates. Among these health challenges, cancer stands out as a critical issue, defined by the unregulated growth of cells in specific organs. The body may develop malignant tumors as a result of this uncontrolled cell division. The most common disease diagnosed globally, lung cancer continues to be the primary cause of cancer-related mortality for both men and women. According to the international agency for research on cancer (IARC) and American cancer society (ACS) Each year, approximately 20 million new cases are identified globally, and over 10 million individuals lose their lives to the disease. Common symptoms of lung cancer include fatigue, unintended weight loss, and hemoptysis (coughing up blood). Other risk factors, including smoking, alcohol use, poor diet, and air pollution, are also linked to lung cancer. Lung cancer can be classified as either small-cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC) based on the histology of the cancer cells. Only 5% of people develop SCLC, while NSCLC accounts for about 85% of all lung cancer cases. Lung cancer has a 5-year survival rate of less than 18%, which is lower than that of other cancer types such as breast cancer (90%), colorectal cancer (65%), and prostate

P. Ranjan et al. (eds.), *Proceedings of the International Conference on Recent Advances in Artificial Intelligence for Sustainable Development (RAISD 2025)*, Advances in Intelligent Systems Research 196,
https://doi.org/10.2991/978-94-6463-787-8_3

than that of other cancer types such as breast cancer (90%), colorectal cancer (65%), and prostate cancer (99%). The medical, biological, and scientific communities need to pay more attention to lung cancer in order to discover new strategies to promote early identification, which aids in medical decision-making and enhances healthcare outcomes. The vast amount of computed tomography (CT) scan image data pertaining to the lungs may be useful for the identification of lung cancer. Machine learning and deep learning algorithms can use these images to improve early cancer diagnosis and prognosis and choose the most effective treatment plan.

Low-dose computed tomography (LDCT) has emerged as a crucial technique for the early detection of lung cancer, especially in high-risk populations. Advances in deep learning techniques have enabled more effective analysis of LDCT images, improving the detection and prognosis of lung cancer. Small lung nodules that might be a sign of early-stage lung cancer can be found with LDCT scans. These scans can be analyzed by deep learning algorithms to find and categorize nodules according to their properties. By combining deep learning with LDCT image analysis and patient clinical data (e.g., age, smoking history, and family history), predictive models can be enhanced. A more thorough risk assessment is provided by this multi-modal approach.

Traditional approaches frequently entail manual interpretation of imaging data, such as chest X-rays and CT scans, by radiologists. This method might be subjective and varies based on the radiologist's experience and ability. LDCT image analysis is automated by deep learning models, namely Convolutional Neural Networks (CNNs). By learning to recognize patterns and characteristics suggestive of lung cancer, these models can greatly improve the accuracy of diagnosis. To provide a more thorough evaluation of lung cancer risk, deep learning models can incorporate multiple data sources, such as imaging and clinical records. The prediction powers are improved by this multi-modal approach. Deep learning (DL) model require large amount and quality dataset for model training.

The objective of this research paper is to make comprehensive analysis of lungs cancer disease prediction. We study machine learning and deep learning model for early prediction of lungs cancer. The challenges and limitations for prediction of lungs disease also discuss in this paper.

2. Deep learning

DL algorithms have simplified the interpretation of complex multi-dimensional data, including images, movies, and multi-dimensional anatomy scans. CNN and the recurrent neural network (RNN) are two well-known deep learning models that are commonly employed for image and sequence data classification. CNN designs frequently include convolutional layers, pooling processes, fully linked layers, and a classification layer.

Changing the weights of the layers that make up the network's architecture is the main focus of a CNN training procedure. This method is considered an NP-hard problem since it is susceptible to multiple local optima. Beyond these local optima, optimization procedures are required. To increase training efficiency and performance, CNNs are taught to optimize techniques such as Adam, Adagrad, AdaDelta, Nesterov accelerated gradient, and stochastic gradient descent (SGD). These techniques are used to adjust the weights and learning rates in order to reduce the losses.

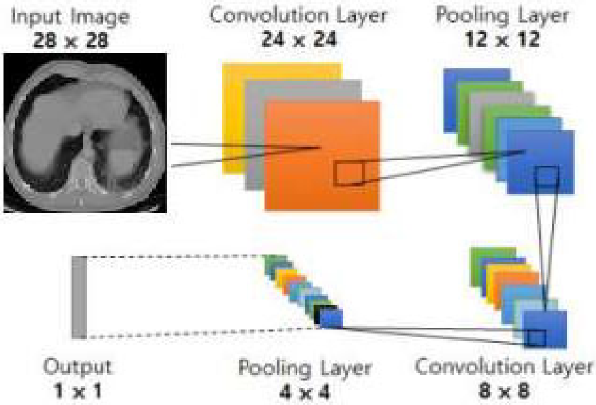

Fig.1. CNN Model

In order to identify lung cancer, Convolutional Neural Networks (CNNs) analyze CT scan pictures by extracting hierarchical features from raw pixel data. First perform the pre-processing of the images. In this phase normalize, resize and enhance the images. In the next phase multiple convolutional layers used to extract features from the images. In the next phase we perform classification of the images using activation functions. The CNN model's operation for an input image is depicted in the above figure. The input image is subjected to a number of filters in the convolutional layer. The input image is filtered, and a feature map is produced. The activation function is implemented after the convolution layer to give the model nonlinearity. The feature map's spatial dimensions are decreased by the usage of pooling layers. This lessens the computing burden and strengthens the model. It connected to the fully connected layer following a number of convolutional and pooling layers. The neural network will provide sophisticated reasoning as a result.

3. Literature review

This section reviews the literature on the efficacy of DL methods for the early diagnosis of lung disease. In order to look into, evaluate, and analyze the consequences of DL on the health sector, we conducted a comprehensive literature analysis. We retrieved research articles from databases like Scopus index journals in order to compile pertinent data.

C. Venkatesh et al. proposed a systematic screening methodology that seeks to rapidly identify lesions associated with lung cancer while enhancing diagnostic accuracy. This methodology utilized Otsu thresholding segmentation to facilitate the precise delineation of the targeted region, and the cuckoo search algorithm was employed to ascertain the optimal parameters for the segmentation of cancerous nodules. The salient features of the lesion were extracted through the application of a local binary pattern technique. A CNN classifier was constructed to ascertain whether a lung lesion is malignant or benign by scrutinizing the extracted features. The proposed framework demonstrates an impressive precision rate of 96.97%. The results of the study suggest that the integration of particle swarm optimization and genetic algorithms has led to improved accuracy and the subsequent synthesis of data [1]. The DNLC model, formulated by H. Elwahsh et al., employs deep neural learning methodologies to forecast oncological conditions. It encompasses a four-step process for the classification of lung neoplasms. Initially, a Deep Network (DN) was utilized to meticulously identify the most advantageous set of features from the available datasets. During the subsequent phase, the authors engaged a deep neural network (DNN) to train utilizing genetic or clinical data repositories.

The third stage involved assessing how well the DNLC model predicted neoplasia in its early stages. Five oncological datasets—colon, lung adenocarcinoma, squamous cell carcinoma, breast, and leukemia malignancies—that have been specially selected for classification purposes are included in DNLC. To predict the performance of the suggested model, experimental evaluations were conducted using the five oncological datasets. The training set, which makes up 80% of the dataset, and the testing set, which makes up 20%, are the two distinct subsets into which the dataset is divided. According to test results, the DNLC framework performs better than alternative approaches in every case, with an average accuracy of 93% [2].

Xinrong Lu et al. employed image processing, deep learning, and metaheuristic techniques to develop the optimal strategy for the early diagnosis of lung cancer. The author of this research developed a novel convolutional neural network. To achieve the optimal layout and increase network accuracy, the marine predator algorithm was also applied. The technique was finally used on the RIDER dataset, and the results were compared to those of other pretrained deep networks, such as CNN ResNet-18, Google Net, Alex Net, and VGG-19. The suggested approach fared better than the other options, according to the final results. With an accuracy of 93.4%, sensitivity of 98.4%, and specificity of 97.1%, the results demonstrated that the MPA-based strategy is more efficient than other state-of-the-art approaches.[3].

To solve the challenge of determining the optimal weight and bias combination for lung cancer classification in CT images, T.I.A. Mohmmad et al. suggested a hybrid approach that combines CNN and metaheuristic algorithms. The author initially created the CNN structure and then identified the solution vector for the model. The ideal weight and bias combination for CNN model training was determined using the EOSA in order to effectively address the classification problem. Significant training of the EOSA-CNN hybrid model allowed the author to find the ideal setup, which produced remarkable results. We used the publicly available lung cancer dataset from IQ-OTH/NCCD [4].

An autoencoder system is used in the method created by Bhoj Raj Pandit et al. to improve overall accuracy and forecast lung cancer. By adding a multispace image to a CNN's pooling layer and optimizing it with the Adam Algorithm, this is achieved. The CT scans were pre-processed using the convolution filter before the image was downsampled using max pooling. Features were then extracted using the autoencoder model, which is based on a convolutional neural network.

In order to reduce inaccuracy during the picture reconstruction process, a multi-space image reconstruction technique was also used. In the end, this results in improved lung nodule prediction accuracy. In order to classify the CT scans, the reconstructed images were ultimately used as input for the SoftMax classifier. The state-of-the-art and suggested techniques were put into practice with Python Tensor Flow. The accuracy of classifying lung cancer increased significantly as a result of this application, rising from 98.9% to 99.5%. Furthermore, the processing time became 12 seconds per second instead of 10 frames per second [5].

A.E. Sebastian and Disha Dua created a novel lung nodule detection algorithm that consists of four steps: segmentation, feature extraction, classification, and image pre-processing. The initial step in this procedure, called pre-processing, was giving the original image a number of different treatments. After that, the previously processed images were segmented using the "Otsu Thresholding model". The next phase involved extracting and classifying the LBP properties using an improved CNN. The activation function and number of convolutional layers of the CNN were adjusted using a developed technique called Improved Moth Flame Optimisation (IMFO). Ultimately, a study based on particular metrics confirmed the efficacy of the approach. The accuracy of the suggested study is significantly higher than that of existing methods. Specifically, its accuracy is 6.85%, 2.91%, 1.75%, 0.73%, 1.83%, and 4.05% higher than that of the SVM, KNN, CNN, MFO, WTEEB, and GWO + FRVM techniques, respectively [7]. Sharmila Nageswaran and her colleagues conducted a study on the precise categorization and prediction of lung cancer using machine learning and image processing technology. First, a set of pictures is needed. The dataset for the experimental study included 83 CT scans from 70 different individuals. When preparing the photos, the geometric mean filter was applied. Consequently, the quality of the images was improved. The K-means technique was then used to separate the images into segments. The segmentation approach can be used to identify the specific region of the image [8].Y.H. Ali et al. created a rigorous fully CNN classifier. The ability to classify biological images was demonstrated by the broad-sense classifier MLCNN. The suggested study employs ML-CNN to identify and classify lung nodules in CT data. Two categories were used to train the proposed ML-CNN classifier: non-nodule (normal) and nodule (malignant, ill, or normal). The sensor data was adjusted before being sent to the ML-CNN with PSO for feature extraction and classification. To lower noise and normalize intensity values, the lung images were pre-processed using a Z-score normalization and a Bayesian cutoff based on a Taylor series.Following an analysis of the color, shape, and intensity of the lung picture, hyperparameters were adjusted using the PSO technique to enhance performance. The ML-CNN with PSO has the following F-measure values: 98.85%, 98.85%, 98.89%, 98.45%, 98.62%, and 98.85%, respectively. In terms of accuracy and convergence, the hybrid approach performs better than the others [9].Supervised learning-trained feed-forward neural networks excel in finding patterns that forecast survival rates in terminal diseases. N.M. Numan et al. used a macro-neural model, which is made up of several basic two-layered neural networks connected in parallel, to tackle the two issues. Macro-neural networks are feed-forward neural networks that have no hidden layers and numerous connected basic networks. The chosen architecture can affect generalization, convergence, and common design issues like as overfitting and hidden layer unknowns. Standard tests were performed on the suggested model. A detailed description of error analysis was given, and multiple data sets were contrasted. In most cases, the suggested approach outperformed traditional methods for predicting survival rates [10]. Vemula Suvarchala et al. created a UCI method to categorize the available data on lung cancer into a malignancy by looking at different machine learning techniques. A well-known Weka classifier approach was used to anticipate and transform the input data into binary form before classifying it as either cancerous or non-cancerous. The comparison technique shows that the proposed RBF classifier was judged the most effective classifier for lung cancer prognostication, with a high accuracy of 81.25 percent [12].

4. Comparative Analysis

This section presents a comparative analysis of the literature review carried out by multiple authors. The following table displays the thorough analysis and the techniques the researchers employed to train the models.

Table 01: Comparative Analysis

Authors and their work	Methodology	Analysis	Limitations
C.Venkatesh and others suggested screening approach [1]	Use Otsu thresholding segmentation	CNN Classifier is used with accuracy of 96.97%. use of Particle swarm optimisation and genetic algorithms resulted in enhanced accuracy	It is sensitive to the noise.
H. Elwahsh and others uses four steps for categorization of lung cancer [2]	Initially use Deep network for analyzing feature from dataset. In the second step use deep neural network to train genetic data. In the third step involved accessing efficacy of DNCL models.	DNLC uses five cancer datasets created especially for classification: blood cancers, breast carcinoma of the squamous cells, a carcinoma of the lungs, and malignancy of the colon. Using the DNLC approach yields an average accuracy of 93%.	Need to improve accuracy of the models.
Xinrong Lu and others Combine image processing, DL, and metaheuristic techniques for enhanced outcomes [3].	They Compared with pretrained deep networks ResNet-18, GoogleNet, AlexNet, VGG-19. RIDER Dataset utilized for testing and validation	MPA was used for optimizing the network layout and Improved accuracy and efficiency. Get accuracy of 93.4%	RIDER dataset has limited number of samples. These models require high computational powers.
The Iraq-Oncology Teaching Hospital's publicly accessible	created a special CNN structure for the first extraction and classification	Examine how well the suggested EOSA-CNN hybrid model performs in	Dataset that may not generalize to other datasets or real-world

IQ-OTH/NCCD Lung Cancer Dataset is used by T.I.A. Muhammad and others [4].	of features. EOSA used to maximize the vector of solutions. To attain the optimal configuration, thoroughly train the hybrid EOSA-CNN model.	comparison to other meta-heuristic strategies and conventional CNN models. 93.21% was attained using the EOSA-CNN model.	situations, but is utilized for evaluation.
Pandit Bhoj Raj and others Compare the suggested autoencoder-based system's performance to current approaches for classifying and predicting lung cancer [5].	preprocessed CT scan pictures using a convolution filter. used an autoencoder based on CNN to extract important characteristics from the processed images.	Enhanced accuracy from 98.9% to 99.5%. The method demonstrated superior accuracy and faster processing, validating its efficiency in lung cancer prediction.	Auto encoders extract features and reduce dimensionality, the resulting data is difficult to read for medical imaging where feature relevance is essential.
The ANN model was developed, tested, and validated using the Survey Lung Cancer dataset by I.A. Nasser and colleagues [6].	Create an artificial neural network (ANN) to detect lung cancer. uses symptoms including coughing, chronic illness, anxiety, yellowing of the fingers, and allergies, among others.	The lung cancer diagnosis accuracy of the ANN model was 96.67%. Using patient data and symptom analysis, the ANN-based method offers a very effective and precise way to diagnose lung cancer.	Lungs cancer symptom may overlap with other respiratory diseases which may lead to incorrect diagnosis.
A.E. Sebastian and Disha Dua Develop an advanced four-stage model for accurate lung nodule detection [7].	Utilized Otsu Thresholding model for precise segmentation of lung regions. Extracted Local Binary Patterns (LBP) features from segmented images.	Features classified using an optimized CNN. Improved Moth Flame Optimization (IMFO) enhanced CNN performance, improving overall classification accuracy.	LBP captures only texture information but fail to get important characteristics of the images.
Sharmila Nageswaran and her colleague collects 83 CT images were collected from 70 unique individuals to form the dataset for the experimental inquiry [8].	The geometric mean filter was utilized during the image preparation phase to enhance image quality.The K-means algorithm was employed to segment the images	ANN, KNN, RF were used for classification. ANN model demonstrated superior accuracy in predicting lung cancer compared to the other algorithms tested.	The image is smoothed using the geometric mean filter, which may result in the loss of edges and tiny details that are essential for precise segmentation in detailed or medical imaging.

To classify biological images, Y.H. Ali and colleagues created a rigorous fully convolutional neural network (ML-CNN) classifier. Two primary categories were used to train the ML-CNN classifier. Nodule: comprises both healthy and cancerous nodules. Normal lung tissue is represented by non-nodules [9].	The lung pictures were subjected to a Bayesian cutoff based on a Taylor series in order to minimize noise. The intensity values in each image were standardized using Z-score normalization. The classifier's performance was improved by adjusting hyper-parameters using the PSO approach.	The ML-CNN with PSO achieved impressive performance metrics Accuracy is 98.45%, Precision is 98.89%, Sensitivity is 98.45%, F-measure is 98.85%	Computational cost is increased by Taylor series extension, particularly for big or high-dimensional datasets.
The goal of N.M. Numan et al. and others is to forecast survival rates in terminal diseases by employing feed-forward neural networks trained by supervised learning. [10]	Several simple two-layered neural networks connected in parallel made up the macro-neural model that was created.	The proposed macro-neural model consistently outperformed standard survival rate prediction methods across most datasets.	This model gives the complex output that having interpretability challenge.
S. Ahmed and others utilized ensemble XgBoost and ResNet101 algorithms to accurately differentiate between NSCLC and SCLC [11].	derived features from the texture GLCM.	With an accuracy rate of 97.0%, KNN beat SVM, which had an accuracy rate of 83.0%.	Feature derived from GLCM not generalised well for different dataset.
Vemula Suvarchala and others Focused on differentiating cancerous from non-cancerous data. Demonstrates the potential of RBF in medical data analysis [12].	Input data pre-processed and converted into binary form. Utilized the Weka classifiers tool for data analysis and classification.	Radial Basis Function (RBF) classifier achieved 81.25% accuracy	For high dimensional features it may face the issue of curse of dimensionality.

The comparative study of the authors' approaches for predicting lung cancer disease is shown in the above table. In general, the aforementioned comparative evaluations have some shortcomings, such as inconsistent model performance, difficulties with interpretability, and a dearth of high-quality datasets to train the models.

5. Conclusion

Lung cancer diagnosis is being rapidly advanced by deep learning technologies. The detailed comparative examination of deep learning models utilized for lung cancer diagnosis is covered in this paper. Successful applications of CNN, RNN, ANN, VGG-16, and SVM for lung cancer detection have been made. The absence of high-quality datasets, model interpretation, and variances in model performance are some of the difficulties. By overcoming present challenges, deep learning has promise for significantly advancing individualized therapy for lung cancer.

Disclosure of Interest: The authors declare that they have no conflict of interest

References

1. Venkatesh, C., Ramana, K., Lakkisetty, S. Y., Band,S. S., Agarwal, S. andMosavi, A.: A Neural Network and Optimization Based Lung Cancer Detection System in CT Images Front. Public Heal. **10**, 1-9 (2022).doi: 10.3389/fpubh.2022.769692.10
2. Elwahsh, H., Tawfeek,M.A.,Abd El-Aziz,A.A., Mahmood,M.A.,Alsabaan,M. and shafeiy,E.E.: A new approach for cancer prediction based on deep neural learning. Journal of King Saud University - Computer and Information Sciences 35(6),(2023).doi: 10.1016/j.jksuci.2023.101565.
3. Lu,X., Nanehkaran,Y.A., and Karimi Fard,M.: A Method for Optimal Detection of Lung Cancer Based on Deep Learning Optimized by Marine Predators Algorithm,Comput. Intell. Neurosci.**2021**(1), 3694723 (2021) doi: 10.1155/2021/3694723.2021,(2021).
4. Mohamed,T.I.A., Oyelade,O.N., and Ezugwu,A.E.: Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm. Plos one **18**.8 (2023) doi: 10.1371/journal.pone.0285796.18
5. Pandit, B. R., et al.: Deep learning neural network for lung cancer classification: enhanced optimization function Multimed. Tools Appl. 82(5), 6605-6624(2023) doi: 10.1007/s11042-022-13566-9
6. Nasser,I. and Abu-Naser,S.: Lung Cancer Detection Using Artificial Neural Network. **3**(3), 17-23(2019)
7. Sebastian, A. E. and Dua, D.: Lung Nodule Detection via Optimized Convolutional Neural Network: Impact of Improved Moth Flame Algorithm. Sens. Imaging. **24**.1 (2023): 11. doi: 10.1007/s11220-022-00406-1
8. Nageswaran,S.,et al.: Lung Cancer Classification and Prediction Using Machine Learning and Image Processing, Biomed Res. Int. **2022**(1), 1755460(2022) doi: 10.1155/2022/1755460
9. HussainAli, Y., et al.: Optimization System Based on Convolutional Neural Network and Internetof Medical Things for Early Diagnosis of Lung Cancer. Bioengineering **10**(3), 320(2023) doi:10.3390/bioeng ineering10030320.
10. Numan, N., Abuelenin,S. and Rashad, M.: Prediction of Lung Cancer Using Artificial Neural Network. Int. J. Intell. Comput. Inf. Sci. **16**(2), 1-19(2018) doi: 10.21608/IJICIS.2018.10013
11. Ahmed, S., et al.: The Deep Learning ResNet101 and Ensemble XGBoost Algorithm with Hyperparameters Optimization Accurately Predict the Lung Cancer. Appl. Artif. Intell. 37(1), 2166222(2023)(2023) doi: 10.1080/08839514

12. Suvarchala, V., Subbareddy, V. and Rao Madala, S.: Lung Cancer Prediction Using Machine Learning Methodologies. Volatiles Essent. Oils **8**(6), 1265–1272(2021)