Intelligent deep learning algorithm for lung cancer detection and classification

N. Sudhir Reddy, V. Khanaa

Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India

Article Info

ABSTRACT

Article history:

Received Aug 18, 2022 Revised Nov 2, 2022 Accepted Nov 27, 2022

Keywords:

Accuracy Convolutional neural network Intelligent deep learning Lung cancer Surgery

Lung cancer is one of the leading causes of cancer mortality. The overlapping of cancer cells makes early diagnosis difficult. When lung cancer is found early, many therapy choices are reduced, the danger of invasive surgery is reduced, and the chance of survival increases. The primary goal of this study work is to identify early-stage lung cancer and categories using an intelligent deep learning algorithm. Following a thorough review of the literature, we discovered that certain classifiers are ineffective while others are almost perfect. In general, several different kinds of images are employed, but computer tomography scanned images are preferable due to their reduced noise. Intelligent deep learning algorithm is one such approach that employs convolutional neural network techniques and has been shown to be the most effective way for medical image processing, lung nodule identification, classification, feature extraction, and lung cancer prediction. The characteristics are taken from the segmented images and classified using intelligent deep learning algorithm. The suggested techniques' performances are assessed based on their accuracy, sensitivity, specificity, recall, and precision.

This is an open access article under the <u>CC BY-SA</u> license.



Corresponding Author:

N. Sudhir Reddy Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research Tiruvanchery, Selaiyur, Chennai, Tamil Nadu 600126, India Email: sudhir.navar@mrce.in

1. **INTRODUCTION**

This coronary artery disease (CAD) based early detection and prediction method takes into account a wide range of key patterns. Other variables that lead to lung cancer include contamination of the environment, including air pollution, and excessive alcohol intake. The chance of acquiring lung cancer is 20-25 times higher for someone who smokes more than one pack of cigarettes each day compared to someone who does not smoke at all. It is believed that lung cancer develops because of out-of-control cell proliferation in one or both of the lungs [1]. If lung cancer spreads to the brain, it may cause visual issues and weakness on one side of the body, among other complications [2]. Figure 1 shows the beginning stages of cancer cells. Symptoms of primary lung cancer include a coughing fit, coughing up blood, chest discomfort, and shortness of breath. Some of the more modern procedures for detecting lung cancer include chest x-rays, computed tomography (CT), magnetic resonance imaging (MRI), and sputum cytology, to name a few examples. These procedures, on the other hand, are out of reach for many people since they are expensive and time-consuming. A new approach for diagnosing lung cancer in its early stages is urgently required since the majority of existing procedures are only capable of identifying lung cancer in its late stages, hence diminishing the patient's probability of surviving the disease. As a result, image processing methods may help to improve the quality of human analysis.



Figure 1. The beginning of cancer

In addition, the use of important pattern prediction techniques in conjunction with this lung cancer risk prediction model will aid in identification procedure. As a result, early prediction should play an important part in the diagnostic process as well as in the development of an effective prevention plan. As previously stated, the present procedures for diagnosing lung cancer are both expensive and time-consuming, making them unaffordable for many people [3]. Additionally, it discovers cancer when it is in an advanced stage, hence decreasing the likelihood of a patient surviving.

As a result, the suggested method is intended to forecast lung cancer in its early stages based on a limited number of indicators and thresholding. There is a reduction in the amount of time and money necessary for excessive medical testing in this system since the number of regulations for testing is decreased. Another benefit is that the suggested system is web-based, allowing patients from different regions to speak with clinicians in real time over the internet. The following are the primary goals of this system: i) to get around certain testing standards that aren't really necessary, ii) to improve the efficiency of the system in terms of time and money, iii) to improve the precision of the system's performance and accuracy, iv) to make use of a smaller number of qualities, and v) to identify at the early stages of the disease.

It is intended to raise the patient's overall survival rate by more than 5 years. The literature survey is carried out in great detail in order to gather as much relevant information as possible on the issue under consideration. The rapid development of AI is fascinating for numerous individuals because of its various applications in different zones. It tends to be utilized for extortion location, PC vision, bioinformatics, and clinical CT image determination. This is utilized for the forecast of malignancy dependent on clinical reports like CT, x-ray, and MRI, and so forth, and it has been demonstrated that due to different AI methods, it has gotten simpler for the specialist to anticipate sickness at the right stage. Malignant growth [4] is a primary reason for death all around the world, and by 2018, it has been assessed as 9.8 million passing; the world wellbeing association has given this assessment, and the most widely recognized malignancy is a cellular breakdown in the lungs. There are different purposes behind malignant lung growth like smoking, traveler to radon gas, and so forth, yet the individual who smokes does not have to experience the ill effects of cellular breakdown in the lungs; it can likewise happen because of used smoking. The treatment is observing and the lung knob investigation [5] by utilizing the CT clinical images Image processing methods are a high-quality tool that may be used to improve the quality of human analysis. Within this suggested approach, the image processing technology is used in order to improve the CT images captured during the early diagnosis and treatment phases of the disease. The major contributions of this work are i) a convolutional neural network (CNN) model is used to extract the detailed features from the lung images and classifies the benign and malignant classes and ii) the simulations performed on LUNA16 dataset shows that the proposed method resulted in better subjective and objective performance as compared to existing approaches discussed in literature.

Rest of this article is contributed as follows: section 2 deals with the literature survey with their drawbacks. Section 3 deals with the analysis of proposed method. Section 4 deals with the results and discussions whereas section 5 deals with conclusion.

2. LITERATURE SURVEY

According to Luna *et al.* [6] explained 'dept unassisted embedding for clustering analysis. Clustering is essential in numerous application fields driven by information and is generally concentrated in distance capabilities and calculation collection. Massively little work has focused on bunching CT images. Any misclassification of any image does not yield an inaccurate result. According to Tuncal *et al.* [7] discussed

about CT based lung image classification. It is a standard methodology to recognize and evaluate cell divides in the lungs. To assess the dangers of lung buttons, clinical practice regularly incorporates master subjective assessments of a few criteria depicting the look and form of a knob, although those highlights are primarily abstract and self-assertive. The model suggested in [8] is clarified that our methodology can be factored into two phases and that each phase may be progressed skillfully throughout deep organizations by way of angle back proliferation. We collect another dataset of 131 neurotic cases that is the most incredible collection for pancreatic sore division, as far as we can tell. Research by Yuan et al. [9] proposed shape-based characteristics were achieved via the application of the Gabor filter. The feature selection process was carried out using the symmetric double sided (SDS) algorithm. It consists primarily of four phases. This work offered a planned cell division in pulmonary recognition to increase precision and yield and reduce found time. Research by Lee et al. [10] proposed a gaussian filter is applied to the input CT image, which aids in the removal of noise and is a particularly successful way of image processing. An additional benefit of the Gaussian filter is that it eliminates unwanted region using marker-controlled watershed segmentation. Research by Nair et al. [11] used logistic regression classifier for identifying the classes of lung cancer. Without aid from the human side, the dice-sorensen coefficient (DSC) calculated by our method to be 63:44 percent average precision, more significant than the number (60:46 percent) without profound supervision. In any event, it offers less accuracy in this interaction. The measurement of information in MR CT images is a lot for manual translation and analysis. Researh by Bartholomai et al. [12] used Naive Bayes classification for lung cancer detection with multiple classes. The pixel is given a value based on whether it is below or over a certain threshold value determined by two levels.

Researh by Han et al. [13] used decision trees classification for lung cancer detection with different classes. A hyperplane is selected in such a way that the margin is maximized. When compared to the other thresholding method, the accuracy is obtained at 100 percent. Researh by Gupta et al. [14] used random forest classification for lung cancer detection. The feature extraction approach evaluates characteristics such as area, perimeter, and eccentricity (roundness) in order to extract useful information. Research by Sim et al. [15] used the texture matching process is then carried out using the local binary pattern (LBP). The performance of LBP is superior than that of other textual patterns that are currently available. The classification process is then carried out using a support vectore machine (SVM) classifier. To reduce noise and improve CT image quality, Pradhan and Chawla [16] used pre-processed utilizing a variety of image improvement methods. Following the conversion of the grayscale CT image for the purpose of image segmentation, further morphological opening procedures were carried out. Researh by Pati [17] used supervised learning classifier (SVM) is used to classify CT images into two categories: normal and abnormal, based on these characteristics. According to the authors, the suggested approach is very accurate in detecting cancer in its early stages. The evaluation of image quality and the improvement of image quality are dependent on the level of enhancement [18]. Preprocessing techniques based on histogram equalization (HE) are used at this step to improve the overall quality of the data. Classification by Yu et al. [19] is particularly significant throughout the digital image analysis process since it categorizes CT images into categories based on their similarities, which is very useful. In the conventional system [20], HE is utilized for preprocessing of CT images, and feature extraction is performed using HE. A CNN classifier [21]–[23] is used to determine if the patient is normal or abnormal, and this is done using HE. Following that, the survival rate of the patient is projected based on the attributes that were retrieved [24], [25]. with practical methodologies to detect malignant lung growth early and screen the seriousness.

3. PROPOSED SYSTEM MODEL

This section comprises of improved dial's loading algorithm (IDLA) procedure utilized for the forecast of malignant growth in both CT image information. That is check report through which we can anticipate the area of tumor or the size of the tumor, and CSV document which contains information like age, sex, smoking rate, and so on. IDLA for lung cancer detection and classification is implemented in four stages namely i) extortion location, ii) PC vision, iii) AI enabled bioinformatics, and iv) clinical CT image determination as shown in Figure 2.



Figure 2. Implementation of IDLA for lung cancer detection

Intelligent deep learning algorithm for lung cancer detection and classification (N. Sudhir Reddy)

This demonstrates how the framework will operate; in this case, the CT filter CT image is first collected from the location with the use of DI-COM software. The scratched information is used to create the dataset, which is then used to pre-process the data. In order to predict the cellular breakdown in the lungs, the datasets are then pre-prepared by switching from dim scale CT images to double CT images and paired CT images. In this encounter, vigilant hash identification is used. Using CNN, these split highlights may be classified according to geography, border, and erraticness. It refers to the total number of pixels in the CT scan of the malignant tumor. The number of 1s in the scalar value is addressed by the deserted area: i) perimeter: it is the genuine number of all pixels which are interconnected on the edges of the tumor, and it is the amount of everyone twofold piece pixels which are available on the diagram of the knob and ii) eccentricity: the roundness or matric worth or anomaly list or circularity is too short of one for other shapes and one for roundabout shape.

Convolutional neural organizations envelop by different layers in their designs. CNN could be a feedforward and amazingly huge methodology, particularly in recognition. Organization structure is fabricated simple; has fewer preparing boundaries. A convolution neural organization includes numerous layers inside the neural organization that comprises one or many convolution layers, thus prevailing by at least one wholly associated layer in different standard layers in the neural organization. Convolution neural organization engineering is ordinarily utilized coordinated effort with the convolution layer and pool layer. Between convolution layers, the pooling layer is visible. It confuses the highlights of the specific position. Since not all the highlights are important, the position and highlights need to be varied. The activities on the pooling layer include max pooling and include pooling. Mean pooling calculates the usual neighborhood inside the element focus, while max-pooling determines the neighborhood within a limited number of highlight areas.

A CNN utilizes the learned highlights with information and utilizes 2D convolutional layers. This infers that this sort of organization is best for handling 2D CT images. Contrasted with different strategies for CT image order, the organization utilizes almost no pre-handling. This implies that they can utilize the channels that clients must work in different calculations. Figure 3 shows CNN's that can be utilized for many applications from image and video recognition, image order, and frameworks for a common language and clinical image analysis.

- Input layer: this layer has the raw pixel upsides of the image.
- Convolutional layer: this layer provides the outcomes of the neuronal layer linked to the input areas. In this layer, we specify the number of filters to be utilized. Each pixel slides across the input data and receives the filters with the highest intensity as the output.
- Rectified linear unit (ReLU) layer: this layer applies to the CT image data an element-wise activation function. We know a CNN is using back propagation. Thus, we use the ReLU function to maintain pixels' equivalent values and not be changed by back propagation.
- Pooling layer: this layer conducts a down sampling operation in width and height along the spatial dimensions resulting in volume.
- Fully connected layer: this layer is utilized to calculate the score classes, i.e., which class has the maximum input numbers.



Figure 3. Proposed IDLA model

4. RESULTS AND DISCUSSION

Throughout the globe, lung cancer is the most common cause of cancer-related mortality. Lung cancer screening using low-dose CT scans is currently being introduced in the United States, and other nations are anticipated to follow suit in the near future. The training dataset obtained after pre-processing of the medical images. This dataset is required for the detection of lung cancer and classification of the same to project the severity of the lung cancer.

4.1. LUNA-16

Many millions of CT scans will be required for CT lung cancer screening, which will place a significant load on radiologists. As a result, there is considerable interest in the development of computer algorithms to improve screening. The discovery of pulmonary nodules on lung cancer screening CT scans is a critical initial step in the study of lung cancer screening CT images, which may or may not reflect early stage lung cancer. For this job, a large number of CAD systems have previously been presented. Automatic nodule identification techniques using the LIDC/IDRI data set will be the primary focus of the LUNA16 challenge, which will be evaluated on a wide scale. The LIDC/IDRI data collection, which includes the annotations of nodules by four radiologists, is made accessible to the public. As a result, the LUNA16 challenge is a totally open challenge. We have tracks for both comprehensive systems for nodule identification and systems that employ a list of likely nodule sites as a starting point for their detection. In addition, we offer this list so that teams may participate using an algorithm that simply estimates the possibility that a certain area on a CT scan has a pulmonary nodule.

4.2. Subjective evaluation

Figures 4(a) and 4(b) shows the preprocessed medical images. Figure 5 depicts a CT scan of the lungs that have been affected by cancer. Furthermore, since a CT scan is loaded with noise from surrounding tissues, bone, and air. It is necessary to pre-process this noise in order for the CAD system to search for the most efficient results.



Figure 4. Medical images (a) input CT images and (b) preprocessed images



Figure 5. Lung cancer image

Figures 6(a)-(c) depicts a small tube in the air conduit system inside the lungs that is a continuation of the bronchi and leads to the alveoli (the air sacs) where oxygen exchange takes place, as seen in the illustration. Bronchiole is the diminutive of bronchus, which comes from the Greek word bronchos, which refers to the bronchial tubes that carry air to and from the lungs. Chronic bronchiolitis is an inflammation of the bronchioles caused by a virus infection, which occurs most often. To account for the chance that some malignant

1752 🗖

development might occur inside the bronchioles (air passages) within the lung, which are seen in Figure 7, this air is incorporated in the finished mask to construct the finalized mask as indicated.



Figure 6. Lung image with disease effected region (a) preprocessed input, (b) ground truth, and (c) segmented outcome



Figure 7. Classified outcome with annotations

4.3. Performance evaluation

This section compares the performance of proposed IDLA approach with the conventional machine learning models. Table 1 compares the performance of proposed IDLA approach with the existing models such as random forest [16], decision tress [17], logistic regression [18], Naive Bayes [19], and SVM [20]. The proposed method extracted the robust features, which resulted in superior performance as compared to existing approaches discussed in literature. Figure 8 shows the graphical representation of performance comparison.

Table 1. Performance comparison						
Method	Accuracy	Sensitivity	Specificity	F1-score	Precision	Recall
Random forest [16]	81.37	80.28	80.92	81.55	80.36	80.84
Decision tress [17]	83.69	81.28	81.05	83.97	80.79	81.00
Logistic regression [18]	84.06	82.74	83.71	86.13	80.94	83.05
Naive Bayes [19]	86.25	85.59	89.10	86.91	85.36	85.20
SVM [20]	88.94	90.12	89.13	90.50	86.00	88.97
Proposed IDLA	92.81	92.85	93.19	93.90	91.88	92.37

Bulletin of Electr Eng & Inf, Vol. 12, No. 3, June 2023: 1747-1754



Figure 8. Graphical representation of performance comparison

5. CONCLUSION

Lung cancer detection is very complicated for physicians. Detection of cancer is treatable in its early stages. The primary time of the cancer prediction system in its early stage is to ensure the patient's treatment is timely. This article proposed a novel algorithm by name IDLA by using CNN model, which combines digital CT image processing and machine learning to identify the cancer cell via IDLAs automatically with minimum iterations. The simulations performed on LUNA16 dataset shows that the proposed method resulted in better subjective and objective performance as compared to existing approaches discussed in literature. This work can be extended to implement with the advanced DLCNN models for improved accuracy.

REFERENCES

- H. Yu, Z. Zhou, and Q. Wang, "Deep learning assisted predict of lung cancer on computed tomography images using the adaptive hierarchical heuristic mathematical model," *IEEE Access*, vol. 8, pp. 86400–86410, 2020, doi: 10.1109/ACCESS.2020.2992645.
- N. Coudray *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018, doi: 10.1038/s41591-018-0177-5.
- Y. Xu *et al.*, "Deep learning predicts lung cancer treatment response from serial medical imaging," *Clinical Cancer Research*, vol. 25, no. 11, pp. 3266–3275, 2019, doi: 10.1158/1078-0432.CCR-18-2495.
- [4] E. S. N. Joshua, M. Chakkravarthy, and D. Bhattacharyya, "An extensive review on lung cancer detection using machine learning techniques: a systematic study," *Journal of critical reviews*, vol. 7, no. 14, pp. 351–359, 2020, doi: 10.31838/jcr.07.14.68.
- [5] D. A. P. Delzell, S. Magnuson, T. Peter, M. Smith, and B. J. Smith, "Machine learning and feature selection methods for disease classification with application to lung cancer screening image data," *Frontiers in Oncology*, vol. 9, pp. 1–8, 2019, doi: 10.3389/fonc.2019.01393.
- [6] J. M. Luna *et al.*, "Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning," *Radiotherapy and Oncology*, vol. 133, pp. 106–112, 2019, doi: 10.1016/j.radonc.2019.01.003.
- [7] K. Tuncal, B. Sekeroglu, and C. Ozkan, "Lung cancer incidence prediction using machine learning algorithms," *Journal of Advances in Information Technology*, vol. 11, no. 2, pp. 91–96, 2020, doi: 10.12720/jait.11.2.91-96.
- [8] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, 2021, doi: 10.1016/j.tranon.2020.100907.
- [9] Q. Yuan et al., "Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer," JAMA Network Open, vol. 4, no. 7, pp. 1–14, 2021, doi: 10.1001/jamanetworkopen.2021.14723.
- [10] S. H. Lee *et al.*, "Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy," *Physics in Medicine & Biology*, vol. 65, no. 19, pp. 1–25, 2020, doi: 10.1088/1361-6560/ab8531.
- [11] J. K. R. Nair *et al.*, "Radiogenomic models using machine learning techniques to predict egfr mutations in non-small cell lung cancer," *Canadian Association of Radiologists Journal*, vol. 72, no. 1, pp. 109–119, Feb. 2021, doi: 10.1177/0846537119899526.
- [12] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2018, pp. 632– 637, doi: 10.1109/ISSPIT.2018.8642753.
- [13] Y. Han et al., "Histologic subtype classification of non-small cell lung cancer using PET/CT images," European Journal of Nuclear Medicine and Molecular Imaging, vol. 48, no. 2, pp. 350–360, 2021, doi: 10.1007/s00259-020-04771-5.
- [14] S. H. Gupta, S. Goel, M. Kumar, A. Rajawat, and B. Singh, "Design of terahertz antenna to detect lung cancer and classify its stages using machine learning," *Optik*, vol. 249, p. 168271, 2022, doi: 10.1016/j.ijleo.2021.168271.
- [15] J. Sim *et al.*, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-67604-3.

- [16] K. Pradhan and P. Chawla, "Medical internet of things using machine learning algorithms for lung cancer detection," *Journal of Management Analytics*, vol. 7, no. 4, pp. 591–623, 2020, doi: 10.1080/23270012.2020.1811789.
- [17] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: an eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2019, doi: 10.1109/ACCESS.2018.2886604.
- [18] S. H. Hyun, M. S. Ahn, Y. W. Koh, and S. J. Lee, "A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer," *Clinical Nuclear Medicine*, vol. 44, no. 12, pp. 956–960, 2019, doi: 10.1097/RLU.00000000002810.
- [19] K.-H. Yu et al., "Reproducible machine learning methods for lung cancer detection using computed tomography images: algorithm development and validation," *Journal of Medical Internet Research*, vol. 22, no. 8, pp. 1–11, 2020, doi: 10.2196/16709.
- [20] L. S. K., S. N. Mohanty, S. K., A. N., and G. Ramirez, "Optimal deep learning model for classification of lung cancer on CT images," *Future Generation Computer Systems*, vol. 92, pp. 374–382, 2019, doi: 10.1016/j.future.2018.10.009.
- [21] N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y.-C. Chen, S.-J. Cheng, and C.-Y. Chen, "Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer," *International Journal of Molecular Sciences*, vol. 22, no. 17, pp. 1–14, 2021, doi: 10.3390/ijms22179254.
- [22] J. A. Alzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna, and C. Thaventhiran, "Boosted neural network ensemble classification for lung cancer disease diagnosis," *Applied Soft Computing*, vol. 80, pp. 579–591, 2019, doi: 10.1016/j.asoc.2019.04.031.
- [23] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.
- [24] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," CA: A Cancer Journal for Clinicians, vol. 70, no. 1, pp. 7–30, 2020, doi: 10.3322/caac.21590.
- [25] Z. Li, J. Shu, B. Yang, Z. Zhang, J. Huang, and Y. Chen, "Emerging non-invasive detection methodologies for lung cancer (review)," Oncology Letters, vol. 19, no. 5, pp. 3389–3399, 2020, doi: 10.3892/ol.2020.11460.

BIOGRAPHIES OF AUTHORS



N. Sudhir Reddy D S S is currently working as Research Scholar in Department of Computer Science and Engineering at Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamil Nadu. He has completed his bachelor's and master's in CSE from JNTU Hyderabad. He is currently working in lung cancer detection. His current research includes deep learning algorithm, cancer detection algorithm, machine learning, and artificial intelligence. He can be contacted at email: sudhir.navar@mrce.in.



Dr. V. Khanaa D S S C is working as Dean Professor in Department of Computer Science and Engineering at Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamil Nadu. He has completed his B.Sc. in Maths from University of Madras, and his master's and Ph.D in 2006 and 2011 respectively from Bharath University, Chennai. He has published more than 200 research articles which includes 151 international journals, 12 national journals and peer-reviewed conference. His area of interest includes bio-molecular computing system, wireless network, neral network, IoT, and deep learning. He can be contacted at email: drvkannan62@bharathuniv.ac.in.