



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 12, December 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Development of Lung Cancer Prediction Framework Based on Enhanced LSTM Network and XGBoost

¹Sukrit Kumar, ²Mr. Deepesh Dewangan

¹M. Tech Student, Department of Computer Science and Engineering, Shri Rawatpura Sarkar University, Raipur, India

²Assistant Professor, Department of Computer Science and Engineering, Shri Rawatpura Sarkar University, Raipur, India

ABSTRACT: Lung cancer is a severe class of cancer disease that begins inside the internal cells of the lungs. Lung cancer has become the most common and intricate illness all across the world. The early detection of lung cancer plays a critical role in the quick diagnosis of patients. Existing research presented distinct Machine Learning (ML) as well as Deep Learning (DL) based approaches for lung cancer analysis. Yet, the developed prediction methods have drawbacks such as less performance in terms of precision, overfitting, etc. This research proposed the development of a lung cancer prediction framework based on enhanced LSTM Network and XGBoost technique. Our proposed framework based on LSTM and XGBoost is validated through the widespread Kaggle datasets for lung cancer. This predictive lung cancer framework processed and analyzed the lung cancer with improved accuracy. Our LSTM and XGBoost model accuracy was found 97.91%. In the future, more advanced research is possible in the field of lung cancer prediction through DL.

KEYWORDS: Deep Learning, Lung Cancer, LSTM, Machine Learning, XGBoost.

I. INTRODUCTION

Lung cancer prediction is a challenging job for the clinicians in modern era. Genetic abnormalities inside the lungs of people lead cells to expand uncontrolled resulting in tumors, which is how lung cancer begins [1]. Although the precise causes of such alterations remain poorly understood, several hazards are recognized to raise the chance of getting lung cancer. A complicated interaction between genetic, external, as well as behavioral variables frequently occurs [2]. The following are a few significant elements linked to the occurrence of lung cancer. The identification as well as categorization of cancers of the lungs have become crucial fields of study in medicine, and the combination of DL as well as ML presents exciting new opportunities [3]. The preliminary diagnosis of lung cancer is frequently difficult for conventional diagnostic techniques, which emphasizes the requirement for cutting-edge technology to improve precision and effectiveness. The identification of lung cancer in its early stages has been made possible by methods of ML-like Support Vector Machines (SVM) and k-nearest Neighbors (KNN) [4]. Such models help identify aberrant patterns that may be signs of possible malignancy by using information taken from clinical imaging data. Yet the ever-changing healthcare market needs systems that can grasp vast facts and their deep linkages [5], [6].

Convolutional neural networks (CNNs) as well as recurrent neural networks (RNNs) in especially, which are part of DL, have demonstrated significant effectiveness in handling the subtleties of lung cancer identification as well as categorization. RNNs are capable of capturing temporal relationships in sequential information, such as time-series records for patients, whereas CNNs are superior at image-rooted studies, successfully identifying patterns as well as architectures in medical pictures. Combining these two structures allows for a comprehensive diagnosis of lung cancer that takes into account both temporal as well as geographical factors [7]. Figure 1 shows the major causes of lung cancer occurrence in humans.

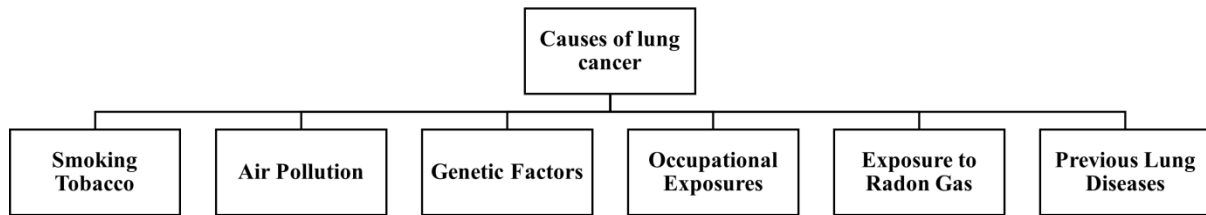


Figure 1: Shows the major causes of lung cancer occurrence in humans.

The ability of ML [8] and DL [9] to identify lung cancer early, which improves prediction as well as treatment results, is one of their main strengths. These models' abilities to learn intricate patterns as well as detect small abnormalities within medical imaging led to improved accuracy and precision in the classification process. The smooth integration of these tools into healthcare facilities is still a difficulty, though. Interpretability, and data privacy, especially the requirement for extensive, varied datasets continue to be important factors. In addition, the current study's endeavors concentrate on the creation of explainable AI models, which aim to illuminate the decision-making procedures of these complex algorithms and foster confidence amongst medical professionals [10], [11].

II. RELATED WORK

The sensitivity, as well as specificity of traditional diagnostic procedures, are frequently limited, which has led to a paradigm change in favor of integrating DL techniques for improved predictive modeling [12]. This overview of the literature examines the development of DL applications in lung cancer prediction, highlighting important techniques, successes, and difficulties. Lung cancer is still a major worldwide health issue, thus attempts to improve early diagnosis as well as prediction modeling must continue [13]. Due to these shortcomings, traditional diagnostic techniques frequently result in late-stage findings with reduced treatment effectiveness. Lung cancer forecasting has advanced in the last several years thanks to the encouraging findings of integrating ML [14] and DL [15] approaches. In [16] A. Elnakib et al. proposed a new CAD system based on DL for the early prediction of lung cancer. To increase the contrasting effect of the small number of images, the suggested approach first preconditions the unprocessed dataset. The next phase is to extract condensed DL features through an investigation of several DL architectures, such as the Alex, and VGG19 models.

M. Sangwan et al. [17] suggested a DL method for diagnosis of lung cancer. Radiological testing is one method of advanced diagnosis that may be employed to identify cancer of the lungs. One technology that may be utilized to analyze lung disorders, such as bronchitis, especially lung cancer, includes chest radiography, which is sometimes known as X-rays. The radiograph picture shows the differences in lung structure among healthy as well as diseased lungs. In [18] R. Pandian et al. proposed a lung cancer prediction model rooted in the CNN as well as Google Net approach. In recent literature, the use of CNNs for lung cancer detection has received a lot of attention. Scholars have investigated the use of several CNN designs for X-ray and computed tomography (CT) scan analysis. Early-stage detection is aided by CNNs' hierarchical feature extraction capabilities, which make it possible to identify tiny irregularities. Research has demonstrated notable progress in terms of sensitivity and specificity, solidifying CNNs' position as a fundamental component of DL-based lung cancer prognosis.

A. Khoirunnisa et al [19] discussed that the use of multimodal data has drawn interest for complete lung cancer estimation, going beyond typical imaging data. To improve prediction models, investigators have looked at combining radionics, genomes, and medical records. Recurrent neural networks (RNNs) as well as multimodal CNNs are two examples of deep learning frameworks that have proven to be adept at discovering intricate associations across a variety of datasets, offering a comprehensive method for predicting lung cancer. S. Mammeri et al. [20] explored that one method to make use of pre-trained models for lung cancer prediction is transfer learning. To efficiently extract characteristics from clinical photos, investigators have modified models that were previously trained on sizable datasets, including ImageNet. By addressing the issues posed by the scarcity of labeled medical datasets, our method shows promise for enhancing generalization as well as accuracy in DL-based lung tumor prediction. Support Vector Machines (SVM) have recently been extensively studied for the categorization of lung cancer, especially in the analysis of imaging data. The capacity of SVM to identify the best hyperplanes towards higher-dimensional environments has demonstrated promise toward the differentiation of benign from malignant lung lesions. To get reliable categorization results, investigators have combined SVM with a variety of kernel features, including the radial basis function (RBF).

III. PROPOSED METHODOLOGY

3.1. Design:

For medical professionals, identifying cancerous lung abnormalities from computerized tomography (CT) images is a challenging yet laborious operation. Computer-aided diagnostic (CAD) solutions have been presented as a way to reduce this strain. DL techniques have demonstrated remarkable achievements in the past few years, surpassing traditional techniques in several domains. These days, scientists are experimenting with various DL methods to improve the effectiveness of CAD tools for CT-based lung cancer detection. Figure 2 illustrates the LSTM and XGBoost framework for lung cancer prediction.

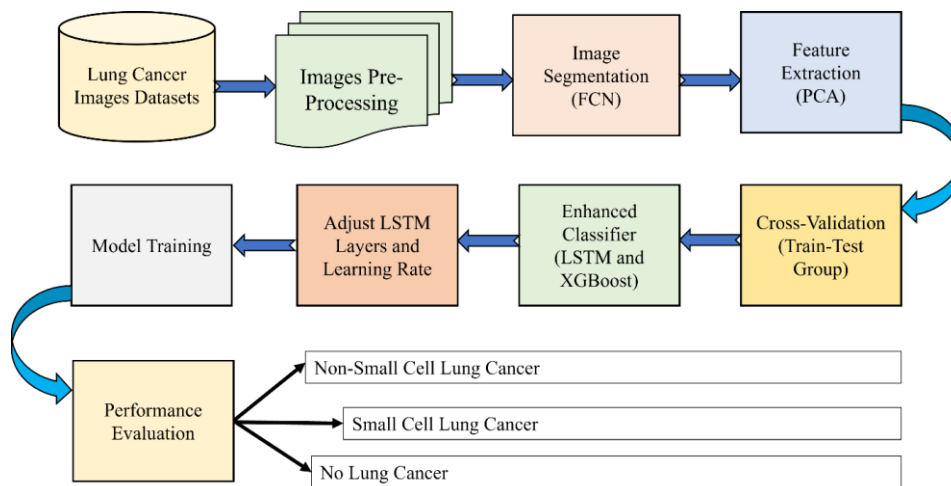


Figure 2: Illustrates the LSTM and XGBoost framework for lung cancer prediction.

This suggested LSTM and XGBoost framework was trained through the well-known dataset Kaggle for lung cancer prediction [21]. Preparing as well as improving clinical imaging datasets is known as "image pre-processing," which is a technique used in conjunction with XGBoost as well as LSTM networks to forecast lung cancer while improving the accuracy of the models. To create an LSTM and XGBoost-based lung cancer forecasting framework, the following image preparation stages are frequently used: (a) collection of dataset and formatting (b) Noise removal, (c) scaling, and (d) data normalization. Further, image segmentation is done, in which FCN (Fully Convolutional Networks) is employed. A computer vision approach called picture segmentation utilizing FCNs divides a photograph into sections that have semantic significance. FCNs are appropriate for segmentation problems since they are made for pixel-wise categorization, in contrast to conventional convolutional neural networks (CNNs), which are made for picture categorization. The feature extraction process is done utilizing the PCA (Principal Component Analysis) approach. A popular method for reducing dimensionality within feature extraction involves PCA. It may be used in computer vision or image processing to minimize the dimensionality of the dataset while keeping the most crucial features. The standard procedure for feature extraction employing PCA is as follows: (a) preparing the data (b) standardization of the data (c) principal components selection. After PCA, the obtained data is split for cross-validation of the model utilizing the train and test data separately. This suggested model utilized 80% lung cancer Kaggle dataset for the training procedure and the remaining 20% dataset was kept for the model validation part. In this proposed model, enhanced LSTM and XGBoost classifiers are utilized for lung cancer prediction. The LSTM network layers and learning rate are adjusted in the model training procedure in real-time. The procedure of training a lung cancer prediction framework with LSTM and XGBoost entails two steps: first, sequential dataset processing for specific input data types (for instance, time series of clinical images or patient records) utilizing the LSTM, and then, ensemble learning for additional forecasting refinement applying XGBoost algorithm. Utilizing the validation set, adjust hyperparameters that include batch size, learning rate, as well as the number of LSTM units to enhance generalization. Analysis of the efficacy of the combined model (LSTM and XGBoost) on the benchmark test set was done. Figure 2 illustrates the common architecture of the LSTM network. For every pattern in the dataset, the proposed combined model retrieves accurate features from the outputs and makes predictions such as no lung cancer, small cell lung cancer, or non-small cell lung cancer.

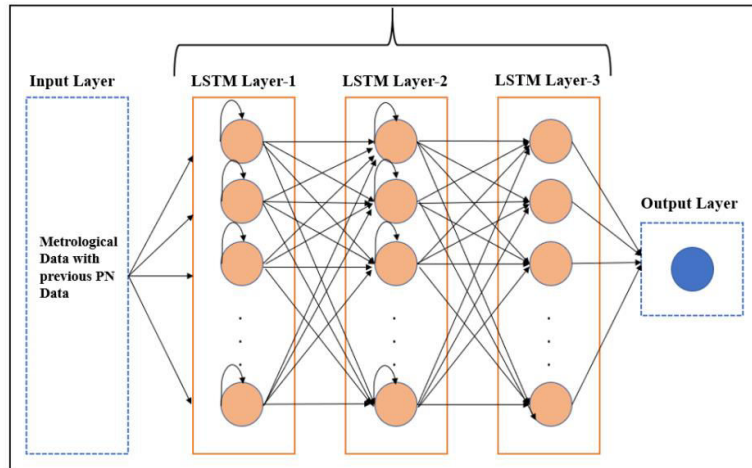


Figure 2: Illustrates the common architecture of the LSTM network [22].

3.2. Pseudo Code:

- Step 1: To include the necessary libraries.
- Step 2: Load as well as preprocess lung cancer datasets.
- Step 3: To segment the images using the FCN approach
- Step 4: To extract the vital features from lung cancer, pre-proceed the dataset by applying the PCA
- Step 5: Split lung cancer data into training-testing sets
- Step 6: Combine the LSTM and XGBoost classifier and adjust the LSTM layers as well as the learning rate
- Step 7: LSTM and XGBoost Model Training initiate
- Step 8: To predict the performance of the proposed lung cancer prediction model

3.3. Instrument:

Parallel processing could be handled by a multi-thread CPU (such as an Intel Core i7) throughout the extraction of features, data preparation, as well as certain aspects of model training. For experiments, a computer system was installed with 16GB RAM, and Windows 7 was chosen for LSTM and XGBoost-based training. Furthermore, lung cancer prediction model implementation is done using Python programming and coding setup is used Google Colab.

3.4. Sample:

Any ML or DL strategy includes datasets for model predictions. The algorithms are developed, trained, and improved by the caliber of the accessible data. For medical imaging applications to be beneficial in any development, the given dataset has to be verified and annotated by specialists. The datasets utilized in current research on deep learning for lung cancer diagnosis are presented in this section. Table 1 shows the lung cancer dataset for proposed LSTM and XGBoost model training.

Table 1: Lung cancer dataset for proposed LSTM and XGBoost model training.

S. No.	Tumor Class	Training sample	Testing sample	Total count
1	Non-small cell lung cancer	800	750	1550
2	Small cell lung cancer	900	650	1550
3	No lung cancer	700	550	1250
5	Overall taken images	2400	1950	4350

3.5. Data Analysis:

Several criteria are utilized to assess how well the created LSTM and XGBoost framework identifies as well as categorizes lung cancer chances in real-time screening. The authors of this research employ statistical metrics, including F1-score, precision, accuracy as well and recall scores. The calculating formula for every metric is defined in the described equations below.

$$Accuracy = \frac{TP+TN}{TS} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{5}$$

In the above formulae, FP shows a False positive, further FN depicts a False Negative, also the definition of TP represents a true positive, as well as TN shows a true negative.

IV. RESULT AND DISCUSSION

Using the synergistic properties of an Improved LSTM network as well as the potent ensemble learning algorithm XGBoost, we set out to construct a highly sophisticated lung cancer forecasting framework in this study. Improving predictability, robustness, as well as interpretability concerning lung cancer forecasting was the aim. We suggest a two-step procedure for our design model. To capture complex temporal connections inside consecutive clinical lung cancer image data and ensure an additional nuanced depiction of the fundamental trends, an Improved LSTM network was first established. The algorithm's ability to extract significant details from the input patterns was further enhanced with the addition of attention algorithms as well as a feature engineering process. After that, a collaborative learning strategy was used, in which the XGBoost technique was used to improve aggregate model accuracy as well as refine forecasts through combining with the LSTM outcome smoothly. A more reliable as well as accurate lung cancer forecasting system was produced by combining the advantages of XGBoost for managing non-linearity as well as feature significance with the sequential modelling capabilities of LSTM.

The efficiency of our suggested framework was shown by the experimental findings obtained from a wide range of datasets. The framework demonstrated its promise for practical clinical implementations by outperforming conventional methods in terms of prediction accuracy. Feature significance assessment improved the model's understanding and offered insightful information about the variables affecting forecasts. Still, there is more work to be done to provide an accurate lung cancer forecast.

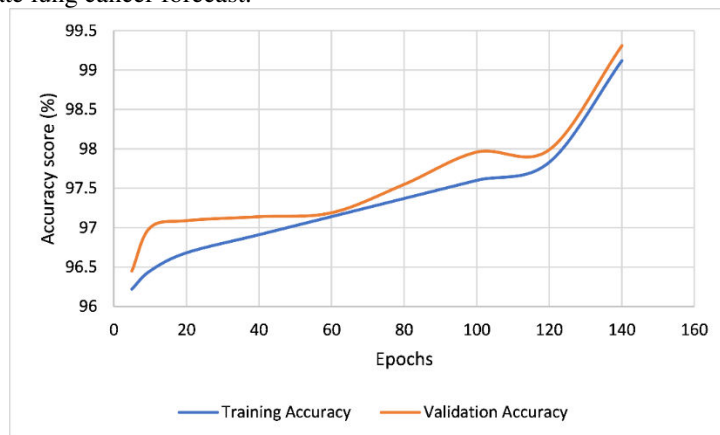


Figure 3: LSTM and XGBoost model accuracy in training and validation.

Figure 3 shows the LSTM and XGBoost model accuracy in training and validation. This LSTM and XGBoost-based model training accuracy on epochs 5, 10 20, 40, 60, 80, 100, 120, and 140, is obtained at 96.22%, 96.45%, 96.68%, 96.91%, 97.14%, 97.37%, 97.6%, 97.83%, and 99.12%. The validation accuracy of LSTM and XGBoost model on epochs 5, 10 20, 40, 60, 80, 100, 120, and 140, is obtained at 96.45%, 97%, 97.09%, 97.14%, 97.19%, 97.55%, 97.96%, 97.99%, and 99.31%.

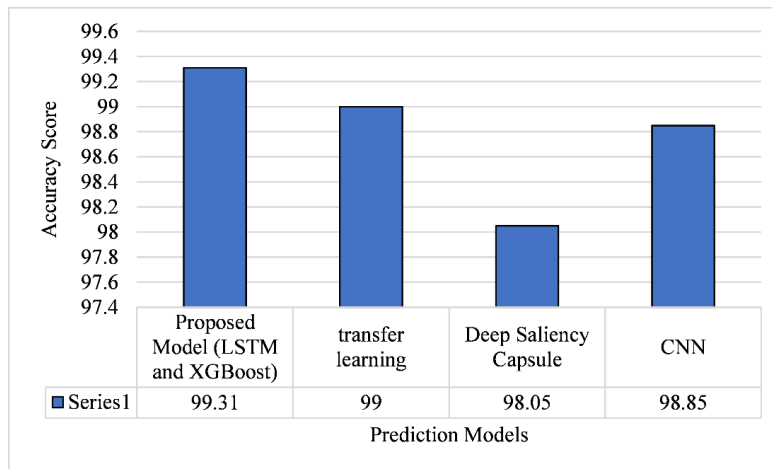


Figure 4: Model accuracy score comparisons [23]–[25].

Figure 4 shows the LSTM and XGBoost model accuracy score comparisons. Our proposed model accuracy is 99.31%. The other models based on transfer learning, deep saliency capsule as well as CNN received less accuracy which is 99%, 98.05%, and 98.85%. The accuracy score of LSTM and XGBoost is the highest in all other techniques.

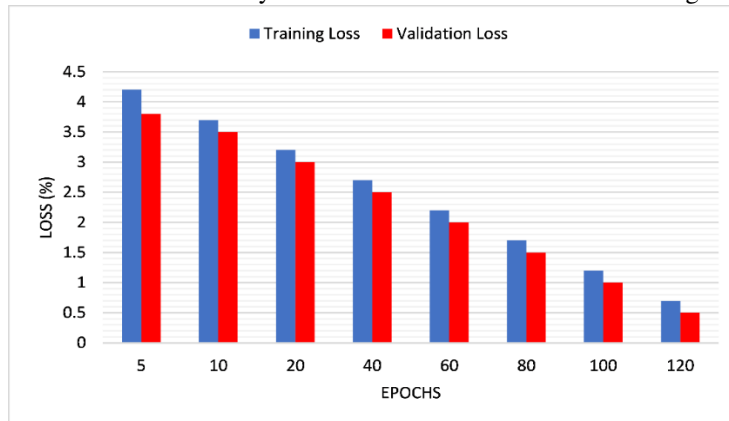


Figure 5: LSTM and XGBoost model loss analysis.

Figure 5 represents our LSTM and XGBoost model loss analysis. The loss score graph clearly shows that in both trains as well as validation of the model, losses are minimal consequently concerning higher accuracy scores on varied epochs. Thus, this proposed approach is the best solution for the early prediction of lung cancer.

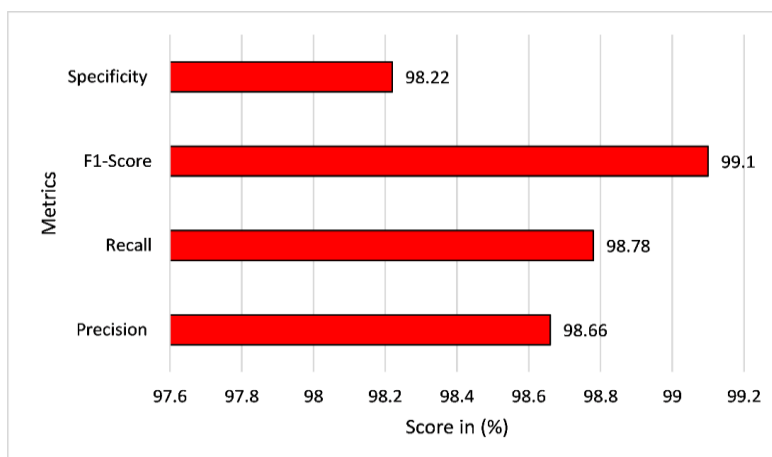


Figure 6: Represents the different metrics analysis for the proposed LSTM and XGBoost Model.

Figure 6 represents the different metrics analysis for the proposed LSTM and XGBoost model. The specificity, F1-score, recall, and precision values are 98.22%, 99.10%, 98.78%, and 98.66%, respectively. Hence, the all-result of our proposed LSTM and XGBoost model is highest in comparison to the earlier methods.

V. CONCLUSION AND FUTURE WORK

Conclusively, our study represents a noteworthy advancement in the creation of an all-encompassing approach for predicting lung cancer by merging the advantages of Upgraded LSTM networks with XGBoost. This suggested approach might completely change the early detection as well as therapy of lung cancer in addition to adding to the expanding corpus of research in healthcare AI. It also shows possibilities for real-world application in clinical environments. Combining ensemble with advanced learning methods is an example of a collaborative method for solving intricate problems in predictive modeling for medical applications. To enhance the prediction capacity of the framework, future research ought to look into the incorporation of other multidimensional data sources, including genetic or radiomics datasets. Furthermore, comprehensive verification on a variety of patient populations as well as cooperation with medical experts would be essential to guarantee the generalizability as well as usefulness of the suggested framework.

REFERENCES

- [1] S. Akila Agnes and J. Anitha, "Automatic lung cancer detection in low-dose lung CTs using transfer learning," *J. Adv. Res. Dyn. Control Syst.*, 2018.
- [2] L. Singh, H. K. Choudhary, S. Singh, A. K. Bisht, P. Jain, and G. Shukla, "Automated Detection of Lung Cancer using Transfer Learning based Deep Learning," 2022. doi: 10.1109/CISES54857.2022.9844372.
- [3] M. Humayun, R. Sujatha, S. N. Almuayqil, and N. Z. Jhanjhi, "A Transfer Learning Approach with a Convolutional Neural Network for the Classification of Lung Carcinoma," *Healthc.*, 2022, doi: 10.3390/healthcare10061058.
- [4] R. AlGhamdi, T. O. Asar, F. Y. Assiri, R. A. Mansouri, and M. Ragab, "Al-Biruni Earth Radius Optimization with Transfer Learning Based Histopathological Image Analysis for Lung and Colon Cancer Detection," *Cancers (Basel)*, 2023, doi: 10.3390/cancers15133300.
- [5] L. T. Omar, J. M. Hussein, L. F. Omer, A. M. Qadir, and M. I. Ghareb, "Lung And Colon Cancer Detection Using Weighted Average Ensemble Transfer Learning," 2023. doi: 10.1109/ISDFS58141.2023.10131836.
- [6] S. U. Atiya, N. V. K. Ramesh, and B. N. K. Reddy, "Classification of non-small cell lung cancers using deep convolutional neural networks," *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-16119-w.
- [7] J. Gao, Q. Jiang, B. Zhou, and D. Chen, "Lung Nodule Detection using Convolutional Neural Networks with Transfer Learning on CT Images," *Comb. Chem. High Throughput Screen.*, 2020, doi: 10.2174/1386207323666200714002459.
- [8] U. Chandran, J. Reys, R. Yang, A. Vachani, F. Maldonado, and I. Kalsekar, "Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care," *Cancer Epidemiol. Biomarkers Prev.*, 2023, doi: 10.1158/1055-9965.EPI-22-0873.
- [9] M. Alameer, H. A. Mengash, R. Marzouk, M. K. Nour, A. M. Hilal, A. Motwakel, A. S. Zamani, and M. Rizwanullah, "Deep Learning Enabled Computer Aided Diagnosis Model for Lung Cancer using Biomedical CT Images," *Comput. Mater. Contin.*, 2022, doi: 10.32604/cmc.2022.027896.
- [10] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," 2018. doi: 10.1109/ICCTCT.2018.8551008.
- [11] I. Nazir, I. U. Haq, S. A. Alqahtani, M. M. Jadoon, and M. Dahshan, "Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion," *J. Sensors*, 2023, doi: 10.1155/2023/6683438.
- [12] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, "Machine learning for early lung cancer identification using routine clinical and laboratory data," *Am. J. Respir. Crit. Care Med.*, 2021, doi: 10.1164/rccm.202007-2791OC.
- [13] S. Nageswaran, G. Arunkumar, A. K. Bisht, S. Mewada, J. N. V. R. S. Kumar, M. Jawarneh, and E. Asenso, "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing," vol. 2022, 2022.
- [14] P. Sandhya Krishna, U. J. Reddy, R. S. M. L. Patibandla, and S. R. Khadherbhi, "Identification of lung cancer stages using efficient machine learning framework," *Journal of Critical Reviews*. 2020. doi: 10.31838/jcr.07.06.68.
- [15] M. A. Thanoon, M. A. Zulkifley, M. A. A. Mohd Zainuri, and S. R. Abdani, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images," *Diagnostics*. 2023. doi: 10.3390/diagnostics13162617.

- [16] A. Elnakib, H. M. Amer, and F. E. Z. Abou-Chadi, "Early lung cancer detection using deep learning optimization," *Int. J. online Biomed. Eng.*, 2020, doi: 10.3991/ijoe.v16i06.13657.
- [17] M. Sangwan, S. Gambhir, and S. Gupta, "Lung cancer detection using deep learning techniques," in *Applying AI-Based IoT Systems to Simulation-Based Information Retrieval*, 2023. doi: 10.4018/978-1-6684-5255-4.ch009.
- [18] R. Pandian, V. Vedanarayanan, D. N. S. Ravi Kumar, and R. Rajakumar, "Detection and classification of lung cancer using CNN and Google net," *Meas. Sensors*, 2022, doi: 10.1016/j.measen.2022.100588.
- [19] A. Khoirunnisa, Adiwijaya, and D. Adytia, "Implementation of CRNN Method for Lung Cancer Detection based on Microarray Data," *Int. J. Informatics Vis.*, 2023, doi: 10.30630/joiv.7.2.1339.
- [20] S. Mammeri, M. Amroune, M. Y. Haouam, I. Bendib, and A. Corrêa Silva, "Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning," *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-16864-y.
- [21] Y. Chen, Y. Wang, F. Hu, L. Feng, T. Zhou, and C. Zheng, "Ldnnet: Towards robust classification of lung nodule and cancer using lung dense neural network," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3068896.
- [22] O. Surakhi, M. A. Zaidan, P. L. Fung, N. H. Motlagh, S. Serhan, M. Alkhanafseh, R. M. Ghoniem, and T. Hussein, "Time-lag selection for time-series forecasting using neural network and heuristic algorithm," *Electron.*, 2021, doi: 10.3390/electronics10202518.
- [23] P. K. Pachala and P. Bojja, "Prediction of Lungs Cancer in Medical Images Using Deep Learning Approach," *Ing. des Syst. d'Information*, 2023, doi: 10.18280/isi.280125.
- [24] K. Ramana, M. R. Kumar, K. Sreenivasulu, T. R. Gadekallu, S. Bhatia, P. Agarwal, and S. M. Idrees, "Early Prediction of Lung Cancers Using Deep Saliency Capsule and Pre-Trained Deep Learning Frameworks," *Front. Oncol.*, 2022, doi: 10.3389/fonc.2022.886739.
- [25] Z. Rustam, S. Hartini, R. Y. Pratama, R. E. Yunus, and R. Hidayat, "Analysis of architecture combining Convolutional Neural Network (CNN) and kernel K-means clustering for lung cancer diagnosis," *Int. J. Adv. Sci. Eng. Inf. Technol.*, 2020, doi: 10.18517/ijaseit.10.3.12113.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor
Impact Factor: 8.379



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 **9940 572 462**  **6381 907 438**  **ijircce@gmail.com**



www.ijircce.com

Scan to save the contact details