Journal of Information Systems Engineering and Management
2025, 10(29s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hybrid Approach to Classify and Detect Lung Cancer using Neural Network by Transfer Learning and Feature Extraction

Mithlesh Arya[1], Amita Gandhi[2], Punam Bhoyar[3], Varun Malik[4], Mudassir Khan[5], Kavita Arora[6]
[1] *Swami Keshvanand Institute of Technology, Management & Gramothan Jaipur, India. mithlesh.arya@skit.ac.in*
[2] *Gateway Institute of Engineering and Technology, Sonipat, India. amitagandhi2010@gmail.com*
[3] *Indira Institute of Management, Pune, India. punam.bhoyar@yahoo.com*
[4] *Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. varun.malik@chitkara.edu.in*
[5] *Department of Computer Science, College of Computer Science, Applied College Tanumah, King Khalid University, Abha, Saudi Arabia. mkmiyob@kku.edu.sa*
[6] *School of Computer Applications, MRIIRS, Faridabad, India. kavita.sca@mriu.edu.in*

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Lung cancer is most common in men in India. Smoking is the main cause of lung cancer, apart from this there are many other reasons also. Lung cancer patients' life can be saved by the early identification. There are many tests to identify cancer of lung like chest X-ray and Computed tomography (CT). In this paper we are using two datasets to validate our work. First dataset-I is numeric data which is collectively based on surveys and second dataset-II contains CT images of lungs. In numeric data we have 15 features and have applied a neural network and ensemble classifier to classify the data into normal and cancerous classes. After that Pearson correlation method is used to select the most prominent features. The best performance is attained when the XGBoost ensemble classifier is used and the obtained values are Precision (95%), Recall (98%) and F1 score (96%). On the second dataset-II we are using CT images dataset and two approaches to classify the data into normal, Benign and malignant classes. Firstly, transfer learning based models VGG16 and VGG19 model are applied on dataset. Dataset is categrioed into three parts like training data, testing data and validation data. Training of model is done with 80% data, then validates with 10% data and finally tested with 10% data. The kappa score and F1 score obtained by VGG16 for lung cancer is 98% and 96.43% and by VGG19 is 97.4% and 96.2%. Second we design CNN for the feature selection and this feature selected methods applied on flatten images. After that random forest, gradient boosting and support vector classifiers have been applied on the features.Then logistic regression method is applied to cateogories the images into three classes such as normal, benign and malignant. Most promienet results are predicted after feature selection and the evalution parameters are F1 score (1.00), precision (1.00) and recall (1.00).This paper enhance the accuracy of lung cancer identification and outperforms traditional computer vision methods across multiple performance metrics.

Keywords: Lung Cancer, Transfer learning, CNN, Feature extraction, XGboost

INTRODUCTION

The cancer of lung continues to be one of the most usual and lethal types of cancer globally [1]. The World Health Organization (WHO) reports that cancer of lung is the leading cause of cancer-related deaths, acounting for over 1.8 million deaths annually. Cancer of lung starts in the lung cells. In figure 1, the image of lung and its cells where lung cancer starts shown.

Early detection and diagnosis are critical for effective treatment and improving patient survival rates. There are multiple methods to diagnose lung cancer like blood test, chest X-ray, Computed Tomography (CT) test and biopsy. The small masses of tissue present in the lung are known as lung nodules. Lung nodules may be of two types: cancerous and non-cancerous, which is also known as malignant and benign. Mainly lung cancer is of two types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC)[2]. NSCLC is most the typical type of cancer of lung and it includes squamous cell carcinoma and adenocarcinoma. Treatment of SCLC is difficult because sometime cancer spreads to other organs of the body. Despite these advancements, the interpretation of medical images still heavily relies on the expertise of radiologists, which can be subject to variability and human error.

This underscores the need for automated, reliable, and accurate diagnostic tools that can assist in lung cancer detection in early stage. The challenge of accurately identifying lung cancer at its early stages is compounded by the complexity and variability of tumor appearance in medical images. Current diagnostic methods, while effective, are

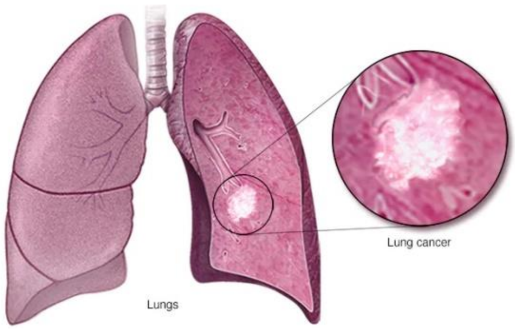

Fig. 1 Lung and Lung cell image

not infallible and can benefit from supplementary automated systems. The development of a robust machine learning model that can accurately classify cancer of lungs images into normal, behign and malignant categories holds promise in addressing this challenge.

LITERATURE REVIEW

Researchers have been proposed many numerous approaches in the literature. Multiple deep learning architectures have been proposed to diagnose the lung cancer in early stages. Pradhan et al. [3] proposed error free system with the help of machine learning and image processing techniques. First CT images of lungs are preprocessed and ROI is finding with the help of segmentation method. In segmentation first images are converted into grayscale images then threshold method is used to convert them into binary images. From the ROI, perimeter, area, eccentricity and GLCM features are extracted. Images are classified into healthy and abnormal with the help of C 4.5 method.

Hussain et al. [4] are using CT image dataset that contain 945 images, 568 images belong to SCLC class and remaining 377 images belong to NSCLC class. Authors used multiple features extraction methods on the dataset and then it was split into training and testing parts. For the classification SVM, Decision Tree and Na¨ıve Bayes methods are used. For the feature extraction they are using texture features, morphological features, RICA features, Auto encoder, Sparse matrix and Entropy. This paper gives best results with RICA and Sparse filter feature extraction method with SVM classifier.

Ahmed et al. [5] proposed a computer aided system using the LDCT dataset. In this dataset 50 images are there. First Deep learning method is used to find the features, so they are using VGG19, VGG16 and Alex for that. After feature extraction, optimal features selection methods are used, they are using genetic algorithms to opt for prominent results. Then in the last step multiple classifiers are used. In their proposed work VGG19 with SVM classifier is giving the best result.

Asuntha et al. [6] used real time CT image dataset for the lung cancer detection. Firstly the images are preprocessed using histogram equalization to improve the quality of image. Then ABC (Artificial Bee Colony) segmentation method is used to get ROI. From this ROI eight different types of features like intensity feature, texture feature, geometric features and volumetric features are extracted using LBP, SIFT and HOG. In the next step prominent features are selected using Fuzzy Particle Swarm Optimization (FPSO). In the last step multiple classifiers like Na¨ıve Bayes, KNN, Adaboost, SVM and ELM are used to categorize the images into healty and cancerous classes. This paper is giving best results using FPSOCNN.

Ruchita et al. [7] have used three datasets for the validation of the work. Authors are using CT images from LIDC-IDRI dataset, Bowl 2017 dataset and LUNA 16 dataset. First preprocessing is done on images using morphological operations. Then CNN model is used for segmentation and feature extraction. They have used combined architecture of U-Net and 3D multipath VGG network for the classification of the lung nodules. N Kalaivani et al. [8] used convolution neural networks for the classification of data into normal and cancerous images. In this 201 CT images are used. CNN with DenseNet and adaptive boosting algorithm is used for the classification of the data.

Smita et al. [9] discussed how multiple machine learning methods are used for the detection of lung cancer. It also discussed Lung cancer detection using IoT devices. A review of lung cancer detection is figured in this paper. In the literature many segmentation, feature extraction and machine learning methods have been used. Researchers have proposed different feature extraction methods [10][11] and machine learning techniques [12][13] for the detection of lung cancer.

Traditional Computer-Aided Detection (CADe) systems for lung nodule detection rely on manually engineered features, which often struggle with early detection accuracy. Convolutional Neural Networks (CNNs) in Deep Learning has significantly improved this field. Jin et al. achieved 87.5% accuracy with a 3D CNN [14], while Shen et al. [15] reached 87.14% using a multi-crop pooling technique. Wang et al.[16] obtained 95.8% sensitivity with a feature pyramid network and a 3D CNN. Sheriff et al.[17] proposed a method for detecting lung cancer by employing the VGG NET 16 deep learning architecture. Lu et al. [18] introduced a dilated Convolutional Neural Network (CNN) based on the VGG16 architecture for lung cancer detection. Raymaha patra et al. [19] explores lung cancer prediction using a CNN with the VGG16 model, highlighting its effectiveness in medical imaging tasks. Thapliyal et al. [20] compare the performance of ResNet50, Xception, and VGG16 models in detecting lung cancer, demonstrating that these AI-based approaches offer efficient and accurate diagnosis. The paper primary endowment is to enhance the accuracy of lung cancer identification and outperforms traditional image processing methods across multiple performance metrics.

1. On numeric dataset-I, we applied neural network and ensemble classifier. Then Pearson Correlation method is applied on the dataset to reduce features.
2. Transfer learning based model VGG16 and VGG19 are applied on the CT image dataset 2. In transfer learning model is already trained with the dataset and starting layers are freeze.
3. In transfer learning few layers are freeze so the values are not updated. So, we applied CNN model on the dataset to extract the features from the flatten layer.
4. Ensemble classifier is used to classify the dataset into three classes.

MATERIAL AND METHODS

Dataset: The methodology for developing a lung cancer detection model involves several crucial steps.Two datasets were taken from kaggle's public available database site. First dataset (dataset -I), Survey Lung Cancer dataset [23] was collected on the basis of patients' survey. It includes 15 features like age, gender, smoking, anxiety, alcohol consumption, allergy, coughing etc. On the basis of these features it is categorized in cancerous and non-cancerous classes. Second dataset (dataset 2), the IQ-OTH/NCCD Lung Cancer Dataset from Kaggle had been selected due to its comprehensive collection of lung CT scan images [21]. This image dataset is essential for the training of the model then the model's performance is evaluated. Dataset 2 was collected from Iraq hospital. This dataset contains 1097 images, 120 images are benign, 561 images are malignant and remaining 416 images belong to normal class. Fig 2 shows the random images from dataset.

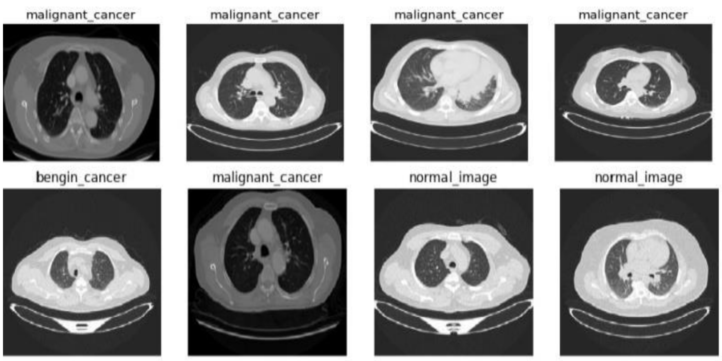

Fig. 2 Random images from dataset 2 with labels

Dataset-I was splited into training data and testing data. Then a sequential classifier was used with an Adam optimizer. Firstly the model use 80% data for training then the model was tested with the remaining dataset.

Precision, Recall and F1 score values are respectively 96%, 94% and 95%. In the next step features are mapped with target class with Pearson correlation. The values are following: GENDER-0.067254, AGE-0.089465, SMOKING-0.058179, YELLOW-FINGERS-0.181339, ANXIETY-0.144947, PEER-PRESSURE-0.186388, CHRONIC DISEASE-0.110891, FATIGUE-0.150673, ALLERGY-0.327766, WHEEZING-0.249300, ALCOHOL CONSUMING-0.288533, COUGHING-0.248570, SHORTNESS OF BREATH-0.060738, SWALLOWING DIFFICULTY-0.259730, CHEST PAIN-0.190451. In this paper top 6 features are selected as the prominent features. Again the model was trained and tested with the 6 features and same ratio. The achived evalution values are 100%, 84% and 91% with respect to Precision, Recall and F1 score.

In the next step, Extreme Gradient Boosting XGBoost ensemble classifier is used on the dataset-I. For this work dataset has been splited into training part and testing part. The model use 80% data for training of model and use remaining 20% data for the testing of the model. This model obtained 95%, 98% and 96% values respectively for Precision, Recall and F1 score. In the next step model has been trained and tested with the 6 prominent features. Now the Precision, Recall and F1 score values are 93%, 95% and 94%. We can see that the value of precision is reduced but the value of recall and F1 score increased.

Neural Network for Dataset-II: On this dataset we have applied two methods for the validation of the work. First transfer learning based VGG16 and VGG19 have been applied on the dataset. In Transfer learning models are already trained with the images so the stating layers are freeze after that we have applied CNN model on the dataset, but there is no significant change in the accuracy. So, we have used CNN for feature selection and then applied random forest, gradient boosting and support vector classifier on the selected features.In the next step logistic regression have been applied on the classifier. There is huge difference in the accuracy.

Transfer Learning: Firstly, the dataset is splited into three parts, namely: training data, testing data, and data for validation in a proportion of 0.8, 0.1, and 0.1, respectively. The portion of traning data confirm that this amount is sufficient for the model training and this willnot overfit or underfit the model. While maintaining data is used for the validation and testing of model and evaluate the model’s performance accurately. Proper data division is critical for preventing overfitting and ensuring that the model generalizes well to new, unseen data. In this paper these steps are used.

1. **Imagedatagenerator:** To enhance the training data’s diversity and improve the model’s robustness, the ImageDataGenerator() class was utilized for data augmentation. Data augmentation helps in creating variations of the existing images, thereby increasing the model’s ability to generalize and perform well on different images.

2. **Visualization:** The image data was visualized to understand the distribution and characteristics of the images in the dataset. This step ensure that the data preprocessing steps were correctly applied and to gain insights into the dataset’s composition.

For the model development, transfer learning with VGG16 and VGG19, a pretrained Convolutional Neural Network (CNN), was employed. The upper layers of the VGG16 and VGG19 model were frozen to retain the learned weights, and only the top layers were retrained on the lung cancer dataset [22]. This approach leverages the pre-trained model’s ability to recognize patterns in images, thereby enhancing the performance on the new dataset with limited data. In VGG19 is little differ from VGG16. The model’s performance is further assessed using precision, recall, F1 score, and Cohen’s kappa. These parameters provide a comprehensive evaluation of the model’s accuracy and reliability in detecting lung cancer from CT scan images.

Kaur et al. (2025): This study presents a high-accuracy method for lung disease classification utilizing logistic regression combined with advanced feature extraction techniques. The research emphasizes the effectiveness of these approaches in enhancing diagnostic precision, showcasing results published in the Egyptian Informatics Journal [23]. Karimullah et al. (2024): This paper introduces an integrated methodology for identification of cancer of lung using CT scans, which incorporates optimization techniques, deep learning algorithms, and IoT data transmission. The findings, published in Frontiers in Oncology, highlight the potential of modern technologies in improving cancer diagnostic processes [24].

CNN and Feature Extraction: In this section, CNN and ensemble learning techniques had been used to develop an image cidentification model to detect cancer of lung into normal, Benign and Malignant classes. These techniques aimed and helped to develop models with enhanced accuracy and robustness. The images are loaded and preprocessed and resized to a consistent size (224x224). CNN architecture was defined to extract features from Fig. 5 Architecture of VGG16 resized images and then perform classification on them. CNN consists of Convolution layers, max pooling layers, flattening layers and dense layers. It is a sequential model built layer by layer. Conv2D is a convolution layer applied on image for the extraction of features from the image. MaxPooling2D is a max-pooling layer applied to reduce the dimension of a feature map. This paper applied Conv2D and MaxPooling2D layes for three times. Then Flatten layer is applied on image to convert the 2D feature map into 1D vector. In the next step Dense layer with relu activation function is used, dense layer is fully connected layer used for classification. After that, a dropout layer is used to prevent overfitting of model. After feature extraction next step is classification of images based on features. Ensemble classifier model is created by staking the base learner models on random forest, gradient boosting and support vector classifiers. CT images are flattened and base learners are trained on the flattened CT images. In this paper logistic regression model is used as meta-learner to combine the base learner predictions. After training, accuracy, recall, precision, and F1-score values are evaluated on the test data.

RESULT & DISCUSSION

The results of the lung cancer detection model demonstrate impressive performance across various metrics. The model achieved an accuracy score of approximately 99.03%, indicating a high level of correctness in classifying the images into the respective categories. This high accuracy reflects the model's capability to distinguish between malignant, benign, and normal cases effectively.

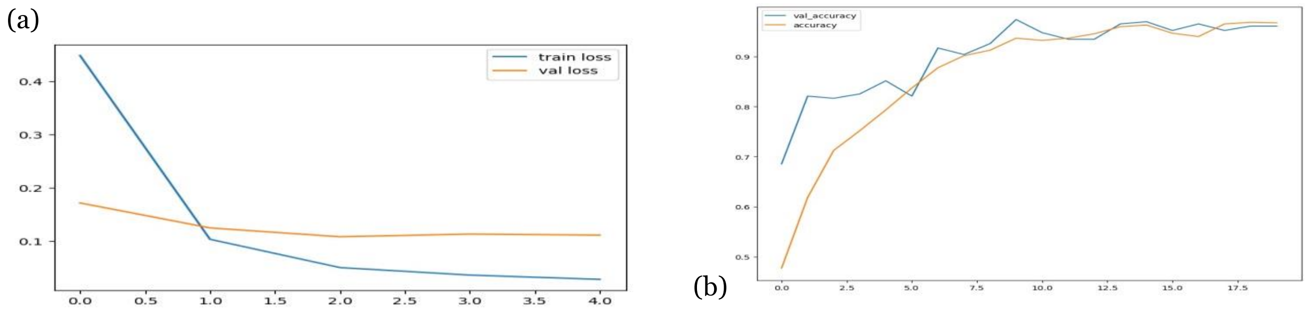

Fig. 3 (a) Train Loss and Validation error curve using VGG16 & (b) Accuracy of CNN and ensemble classifier

As the number of epochs increases, both the training and validation loss decrease, demonstrating that the algorithm is improving its performance over time. This trend signifies that the model is effectively learning from the data, enhancing its ability to generalize well on unseen data. The continuous minimization of loss with increasing epochs confirms the robustness and reliability of the model in accurately detecting lung cancer from CT scan images. The F1 score, which measures the balance between precision and recall, was reported at 96.43%. This high F1 score indicate that the model performs well in both identifying positive cases (malignant and benign) and minimizing false positives, making it reliable for practical applications in medical imaging. The Cohen's kappa score of 98.02% reflects an almost perfect agreement between the model's predictions and the actual labels. This metric shows that the model's performance is robust and consistent, minimizing the likelihood of classification errors.

Table 2 Performance measure of model

Model	Accuracy	F1-Score	Precision	Recall	Kappa		
VGG16	99.083	96.43	98.94	94.44	98.02		
VGG19	98.91	96.2	97.99	94.44	97.04		
Before Feature Selection				**Before Feature Selection**			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
Benign	0.99	0.97	0.98	1.00	1.00	1.00	416
Malignant	0.98	1.00	0.99	1.00	1.00	1.00	608
Normal	0.99	0.97	0.98	1.00	1.00	1.00	120

We have applied an ensemble classifier before and after selection of relevant features. The obtained accuracy using logistic regression model is 0.99 and precision, recall and F1-score values are shown in Table 1. After feature selection the obtained accuracy is 1.00. Moreover, after feature selection the accuracy of the model has been increased. Figure 3 shows the accuracy graph for the proposed models.

Overall, the results suggest that the model performs exceptionally well in detecting lung cancer from CT scan images, with high accuracy, precision, recall, and agreement scores. This performance underscores the effectiveness of the transfer learning approach with VGG16 and VGG19. With the help of feature selection and ensemble learning the performance of the CT image lung cancer have been improved. These results confirm the effectiveness of using transfer learning for lung cancer detection and highlight the model's capability to accurately and reliably classify lung cancer images.

CONCLUSION

This paper successfully developed a lung cancer image classification model using CNN and ensemble learning techniques. For this work two publicly available datasets are used. First dataset is based on a survey. On this dataset Neural network and ensemble learning methods have been implemented and in the next step relevant features are selected with the help of Pearson correlation method. For this dataset, relevant features are not working well. On second dataset lung cancer detection model was developed using transfer learning with the VGG16 and VGG19 architecture. By leveraging a dataset from Kaggle, the model was trained, validated, and tested with a well-structured dataset split of 80% for training, 10% for validation, and 10% for testing. The approach involved freezing the upper layers of the VGG16 and VGG19 network to retain pre-trained features while fine-tuning the model on the specific dataset. The model achieved an exceptional performance with an accuracy of 99.03% and 98.9% indicates a near-perfect agreement between the model's predictions and the actual outcomes. Meanwhile, CNN model extracted features from the CT scan images, while the ensemble learning approach combined the predictions of multiple base learners to improve accuracy and robustness. The final stacking classifier achieved satisfactory performance, demonstrating the effectiveness of the combined approach for medical image classification. For the second dataset, our proposed model obtained more accurate results.

REFERENCES

[1] www.my.clevelandclinic.org/health/diseases/4375-lung-cancer.

[2] https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptomscauses/syc-20374620.

[3] Pradhan, Kanchan, and Priyanka Chawla. "Medical Internet of things using machine learning algorithms for lung cancer detection." Journal of Management Analytics 7, no. 4 (2020): 591-623.

[4] Hussain, Lal, Majid Saeed Almaraashi, Wajid Aziz, Nazneen Habib, and Saif-Ur-Rehman Saif Abbasi. "Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features." Waves in Random and Complex Media 34, no. 1 (2024): 226-251.

[5] Elnakib, Ahmed, Hanan M. Amer, and Fatma EZ Abou-Chadi. "Early lung cancer detection using deep learning optimization." (2020): 82-94.

[6] Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." Multimedia Tools and Applications 79, no. 11 (2020): 7731-7762.

[7] Tekade, Ruchita, and K. Rajeswari. "Lung cancer detection and classification using deep learning." In 2018 fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-5. IEEE, 2018.

[8] Kalaivani, N., N. Manimaran, S. Sophia, and D. D. Devi. "Deep learning based lung cancer detection and classification." In IOP conference series: materials science and engineering, vol. 994, no. 1, p. 012026. IOP Publishing, 2020.

[9] Raut, Smita, Shraddha Patil, and Gopichand. "Lung cancer detection using machine learning approach." International Journal of Advance Scientific Research and Engineering Trends (IJASRET) (2021).

[10] Li, Liangyu, Jing Yang, Lip Yee Por, Mohammad Shahbaz Khan, Rim Hamdaoui, Lal Hussain, Zahoor Iqbal et al. "Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques." Heliyon 10, no. 4 (2024).

[11] Maleki, Negar, and Seyed Taghi Akhavan Niaki. "An intelligent algorithm for lung cancer diagnosis using extracted features from Computerized Tomography images." Healthcare Analytics 3 (2023): 100150.

[12] Yamini, B., K. Sudha, M. Nalini, G. Kavitha, R. Siva Subramanian, and R. Sugumar. "Predictive modelling for lung cancer detection using machine learning techniques." In 2023 8th International Conference on Communication and Electronics Systems (ICCES), pp. 1220-1226. IEEE, 2023.

[13] Wankhade, Shalini, and S. Vigneshwari. "A novel hybrid deep learning method for early detection of lung cancer using neural networks." Healthcare Analytics 3 (2023): 100195.

[14] Jin, Taolin, Hui Cui, Shan Zeng, and Xiuying Wang. "Learning deep spatial lung features by 3D convolutional neural network for early cancer detection." In 2017 international conference on digital image computing: techniques and applications (DICTA), pp. 1-6. IEEE, 2017.
[15] Shen, Wei, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification." Pattern Recognition 61 (2017): 663-673.
[16] Wang, Tonghe, Yang Lei, Sibo Tian, Tian Liu, Walter J. Curran, Kristin Higgins, and Xiaofeng Yang. "Lung tumor segmentation of PET/CT using dual pyramid mask R-CNN." In Medical Imaging 2021: Image Processing, vol. 11596, pp. 841-846. SPIE, 2021.
[17] Sheriff, S. Thanzeem Mohamed, J. Venkat Kumar, S. Vigneshwaran, Aida Jones, and Jose Anand. "Lung cancer detection using VGG NET 16 architecture." In Journal of Physics: Conference Series, vol. 2040, no. 1, p. 012001. IOP Publishing, 2021.
[18] Lu, Yu, Huanwen Liang, Shijie Shi, and Xianghua Fu. "Lung cancer detection using a dilated CNN with VGG16." In Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning, pp. 45-51. 2021.
[19] Raymahapatra, Prithwish, Avijit Kumar Chaudhuri, Sulekha Das, and Alex Khang. "Lung Cancer Prediction Using Convolutional Neural Network (CNN) with VGG16 Model." In Computer Vision and AI-Integrated IoT Technologies in the Medical Ecosystem, pp. 244-254. CRC Press.2023.
[20] Thapliyal, Nitin, Manika Manwal, Vinay Kukreja, and Rishabh Sharma. "Artificial Intelligence-Based ResNet50, Xception, and VGG16 Models for an Efficient Detection of Lung Cancer." In 2024 5th International Conference for Emerging Technology (INCET), pp. 1-5. IEEE, 2024.
[21] Alyasriy, hamdalla; AL-Huseiny, Muayed (2023), "The IQ-OTH/NCCD lung cancer dataset", Mendeley Data, V4, doi: 10.17632/bhmdr45bh2.4.
[22] Ter-Sarkisov, Alex. (2020). Network of Steel: Neural Font Style Transfer from Heavy Metal to Corporate Logos. 10.5220/0009343906210629.
[23] https://www.kaggle.com/datasets/ajisofyan/survey-lung-cancer
[24] Swapandeep Kaur, Sheifali Gupta, Deepali Gupta, Sapna Juneja, Ali Nauman, Mudassir Khan, Izhar Husain, Asharul Islam, Saurav Mallik, High-accuracy lung disease classification via logistic regression and advanced feature extraction techniques, Egyptian Informatics Journal, Volume 29, 2025, https://doi.org/10.1016/j.eij.2024.100596
[25] Karimullah S, Khan M, Shaik F, Alabduallah B and Almjally A (2024) An integrated method for detecting lung cancer via CT scanning via optimization, deep learning, and IoT data transmission. Front. Oncol. 14:1435041. doi: 10.3389/fonc.2024.1435041