Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Hybrid Approach of Multiclassification for Lung Cancer Types Identification

Ramesh Vaishya[1,*], Praveen Kumar Shukla[1]

[1] Department of Computer Science, BBD University, Lucknow, Uttar Pradesh 226028, India

ARTICLE INFO

Article history:
Received 21 January 2024
Received in revised form 7 March 2024
Accepted 28 August 2024
Available online 1 October 2024

Keywords:
Computed Tomography (CT) scan, VGG16, Lung Cancer, Machine Learning

ABSTRACT

As a widespread health concern, lung cancer requires sophisticated detection techniques for a precise classification. This research proposes an integrated framework for improving how well lung cancer is detected by combining processing methods for images and machine learning. The dataset, which includes normal cases, large-cell carcinoma, squamous cell carcinoma, and adenocarcinoma, allows for a detailed investigation of various forms of lung cancer. K-means clustering is the initial stage of the approach of segmenting complex images. Utilizing the VGG16 model, features are extracted that are important for classification. Several classification models, such as SVM, Logistic Regression, and Feedforwa w rd Neural Network, are trained using the combined features from clustering and VGG16. This research goes beyond the dichotomy of benign and malignant cases, investigating in more detail the subtypes of large-cell carcinoma, squamous cell carcinoma, and adenocarcinoma. Including different classes of lung cancer increases the granularity of the classification process, improving diagnostic accuracy and clinical insights. The results highlight how well the suggested strategy works, which is a major advancement in the accurate identification of distinct kinds of lung cancer.

1. Introduction

In the area of world health, lung cancer is a deadly disease that takes millions of lives every year estimated 422 lives are lost every day worldwide [10]. A majority of lung cancer cases are detected in people who are older than 50 years of age. The incidence of lung cancer is increasing every day [11]. It demands cutting-edge methods for early identification and categorization. In the ever-changing field of medical technology, combining advanced techniques like image processing and machine learning presents a viable way to improve the precision and granularity of lung cancer diagnosis [17]. This study aims to further the current discussion by putting forth a thorough framework that explores the complex categorization of individual lung cancer types in addition to making the distinction between cancerous and non-cancerous cases.

* *Corresponding author.*
E-mail address: ramesh.rv.lko@gmail.com

https://doi.org/10.37934/araset.52.1.132150

With great care selected to represent the wide range of lung cancer types, including adenocarcinoma, large-cell carcinoma, squamous cell carcinoma, and normal cases, is the dataset that is being examined. With the help of this large dataset, a nuanced investigation of the different types of lung cancer can be conducted, illuminating the complex manifestations that require specialized diagnostic techniques [18].

The diagnosis and treatment of lung cancer pose distinct challenges due to its ongoing complexity and multifaceted nature. With the growing use of medical imaging, especially computed tomography (CT) scans, there is a chance to better understand and detect lung cancer by utilizing cutting-edge technologies. Because CT scans are so good at finding even the smallest lesions, they are essential for diagnosing lung cancer [12]. This work seeks to break through traditional diagnostic paradigms and usher in a new era of nuanced classification. It is based on a dedication to precision medicine. Advanced X-ray technology is used during a CT scan to take pictures of the human body from several perspectives. Subsequently, the photos are loaded into a computer, which then uses its processing capacity to produce an interior organ and tissue cross-section picture [9].

Numerous methods have been devised to compute handcrafted aspects, including the form of nodules [13–15,40]. Studies on the use of CAD in the detection of lung cancer have employed machine learning to distinguish between benign and malignant nodules. It is challenging for a medical professional or radiologist to promptly and accurately identify cancer because of the volume of CT images. But thanks to technological advancements, Computer-Aided Diagnosis (CAD) can be used to finish this task quickly and effectively. This procedure consists of two distinct steps: the first identifies every nodule visible on the CT scan, and the second classifies any lung nodules that are found. Generally speaking, a CAD system consists of the steps listed below in Figure 1.

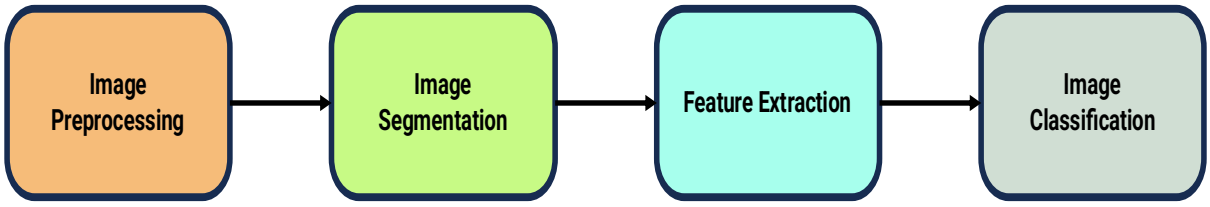


Fig. 1. Basic steps involved in a CAD system

This research is based on the understanding that lung cancer is not a single, homogenous disease, but rather that it is composed of several histological subtypes, each of which has unique traits and clinical implications [20]. With its origins in the lung's glandular cells, adenocarcinoma is the most common subtype and accounts for a large percentage of non-small cell lung cancers (NSCLC). However, large-cell carcinoma presents unique difficulties in both diagnosis and treatment due to its fast growth and tendency to metastasize. The identification of squamous cell carcinoma, which is frequently associated with smoking, requires specific methods because it presents centrally in the larger airways of the lung [20].

The normal class included in the dataset serves as a crucial point of comparison for distinguishing between pathological and healthy lung scans. This large dataset, which accurately captures the complexity of lung cancer in real life, serves as the basis for this study. Along with admitting the intrinsic variability of lung cancer, it also lays the groundwork for developing a dependable and extensively applicable classification model.

The research's methodological approach is in line with the complexities involved in classifying lung cancer. Regions of interest within CT scans are identified using K-means

clustering, a mainstay of image segmentation [22]. It is intentionally chosen to cluster in order to decipher the spatial distribution of pathological findings in the lung, which may not follow well-established anatomical boundaries which is a crucial stage in removing relevant information from the complicated structure of CT scans. The VGG16 model is a popular choice for high-level feature extraction because of its deep convolutional architecture and pre-trained weights on ImageNet [16,23]. By using the knowledge encoded in the VGG16 model from various image datasets, this transfer learning strategy makes sure that the features extracted are not limited by the peculiarities of particular lung cancer dataset [23].

Nevertheless what really makes this research unique is its unwavering dedication to the intricate categorization of various forms of lung cancer. Although binary classifications are useful, a more detailed approach is necessary due to the realities of clinical practice. Squamous cell carcinoma, large cell carcinoma, and adenocarcinoma are the three most common kinds of cancer. Figure 2-4 displays CT scan pictures of patients with large-cell carcinoma, squamous cell carcinoma, and adenocarcinoma, respectively and for just purpose a normal person not diagnosed with any cancer type has been shown in Figure 5.

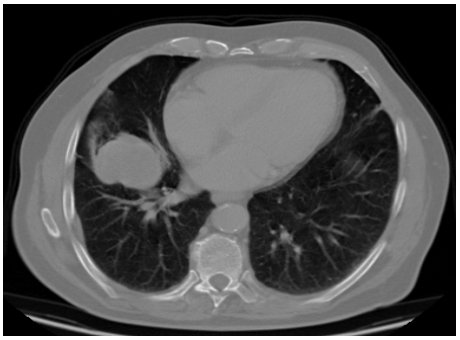


Fig. 2. Adenocarcinoma

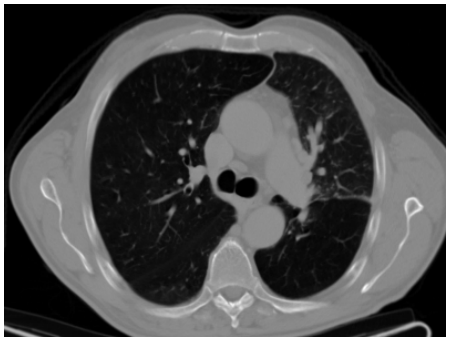


Fig. 3. Large cell Carcinoma

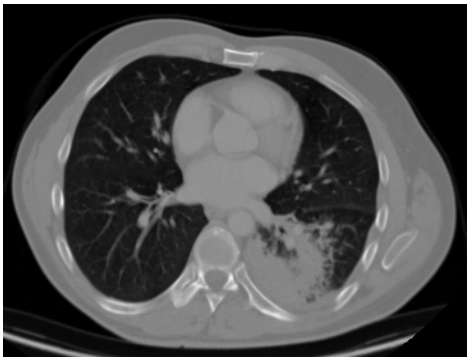


Fig. 4. Squamous cell carcinoma

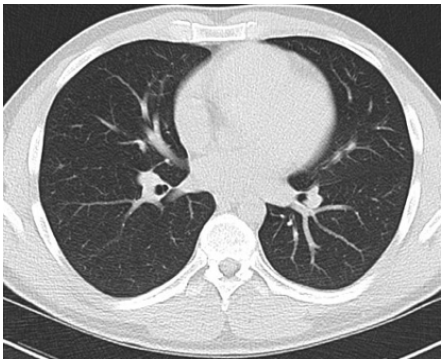


Fig. 5. Normal

Each has its own set of problems and effects. This hierarchical classification recognizes the complex characteristics of lung cancer and offers physicians information that goes beyond a straightforward categorization of lung cancer as either malignant or non-cancerous.

The integration of various classification models, each of which contributes to the overall objective of accurate lung cancer detection, represents the research's conclusion. Three strong contenders stand out: feedforward neural networks, logistic regression, and support vector machines. Each provides a different viewpoint on the dataset. The evaluation metrics offer a thorough understanding of each model's performance, comprising F1 score, recall, accuracy, and precision, in identifying and categorizing the kind of lung cancer, we attained a high accuracy of 99.33%.

This research imagines a future where cutting-edge technologies seamlessly integrate into clinical workflows, providing clinicians with a nuanced understanding of the diseases they seek to diagnose and treat, as it navigates the complex landscape of lung cancer diagnostics. Our dedication to precision medicine—where treatments are customized to each patient's specific needs—is demonstrated by the investigation of various forms of lung cancer. By conducting this research, we hope to significantly contribute to the ongoing efforts to reshape the field of cancer diagnosis and treatment in addition to furthering our scientific understanding of lung cancer.

1.1 Literature Survey

Various methods for image pre-processing, segmentation, and feature extraction have been investigated in the literature currently in publication with regard to the diagnosis of lung cancer. Researchers' various approaches are compiled in the following Tables 1-3:

Table 1
Comparing various image pre-processing methods

Reference	Methods	Uses
Palani and Venkatalakshmi [1]	Gaussian and Gabor filtering	Blurring for noise reduction, Gabor in order to analyze texture
Geng *et al.,* [3]	CLAHE	Enhancement of local contrast
Saini *et al.,* [4]	Wiener filtering	Linear time-invariant filtering for noise removal
Nageswaran *et al.,* [2]	Median filtering	Removal of salt & pepper noise
Nithila and Kumar [5]	Adaptive bilateral filtering	Sharpness enhancement & noise removal
Sangamithraa and Govindaraju [6]	Gaussian and Convolutional filtering	Enhancement of local discontinuities

Table 2
Comparing different image feature extraction methods

Reference	Method of Feature Extraction	Aspects/Features Taken into Account
Palani and Venkatalakshmi [1]	CNN, ARM, DT	Various features for predictive modeling
Nageswaran *et al.,* [2]	RF, KNN, ANN	Noise reduction, feature extraction, damage region analysis
Geng *et al.,* [3]	Multi-Layer Perceptron (MLP)	Pixel-by-pixel predictions for precise segmentation
Saini *et al.,* [4]	Not specified	Examination of digital dental X-ray images
Nithila and Kumar [5]	SBGF-new SPF function	Accurate lung segmentation with selective Gaussian and binary filters
Sangamithraa and Govindaraju [6]	Fuzzy EK-mean segmentation	Contrast, homogeneity, area, correlation, entropy
Lanjewar *et al.,* [34]	DenseNet201, Feature selection methods	Precise feature extraction

Table 3
Comparison of models used

Reference	Models Used
Palani and Venkatalakshmi [1]	CNN, ARM, DT
Nageswaran *et al.,* [2]	RF, KNN, ANN
Geng *et al.,* [3]	VGG-16, Multi-Layer Perceptron (MLP)
Saini *et al.,* [4]	Not specified
Nithila and Kumar [5]	Active contouring, SBGF-new SPF function
Sangamithraa and Govindaraju [6]	EK-mean clustering, Fuzzy EK-mean segmentation, Neural network classification
Lanjewar *et al.,* [34]	Modified DenseNet201, Various machine learning classifiers

A predictive modeling method was presented by Palani and Venkatalakshmi [1] for ongoing observation of lung cancer patients. To achieve continuous monitoring, the methodology combined categorization with a fuzzy cluster-linked augmentation. The application of fuzzy clustering, which is essential for precise picture segmentation, was a fundamental component of this strategy. In order to distinguish between features unique to the lung cancer image and those of the transitional region, the researchers chose to employ the fuzzy C-means clustering technique. Notably, the Otsu thresholding technique was used to distinguish the transition area from the lung cancer representation. The study combined a morphological thinning procedure with the right edge picture to further enhance segmentation. The final goal of the research was to apply convolutional neural networks (CNN), association rule mining (ARM), and classic decision trees (DT). These approaches were combined with a novel incremental classification strategy to produce accurate and incremental classification. The latest health data gathered from IoT devices connected to patients and standard images from the database were essential to the operations' successful completion. The definitive results showed a noteworthy enhancement in the predictive modeling system's accuracy for lung cancer.

Nageswaran *et al.*, [2] used machine learning and image processing to classify lung cancer in their groundbreaking work. Their approach included a detailed review of the patients' medical history in addition to noise reduction, feature extraction, and damage region analysis. The dataset that served as the basis for the study included 83 CT scans from 70 different people that had been carefully preprocessed with a geometric mean filter to enhance the quality of the images. Image segmentation was able to be done accurately by using the K-means algorithm, allowing lung cancer affected areas to be identified. Most importantly, the group used machine learning methods like Random Forests (RF), k-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). Notably, their results highlighted how well the ANN model predicted lung cancer, which represents a significant advancement in the combination of image processing and machine learning technologies.

Geng *et al.,*'s method [3] of combining VGG-16 architecture with dilated convolution offers a novel way to precisely segment lung parenchyma. The approach uses the first three components of the VGG-16 network to conduct input image pooling and convolutions. Next, multiple sets of dilated convolutions are applied to the network to ensure that its receptive field is large. The features of the multi-scale convolution are combined in the last stage, and a Multi-Layer Perceptron (MLP) is used to make pixel-by-pixel predictions in order to precisely segment the parenchymal region. Experimentation evaluation on 137 images shows notable improvements over the state-of-the-art techniques, with a noteworthy Dice similarity coefficient (DSC) metric of 0.9867. The results affirm the effectiveness of the proposed

method in accurately segmenting the lung parenchymal area, outperforming conventional techniques.

Saini *et al.,*'s research [4] offers insightful information about the difficulties in identifying lung cancer and highlights the importance of early detection for a disease with high incidence and fatality rates. Since lung cancer is among the deadliest types of cancer, the study aims to improve the level of analysis provided by digital dental X-ray images by addressing image noise. The most reliable method for detecting lung cancer in clinics is still pathology diagnosis, despite continuous efforts to develop diagnostic instruments. The identification of lung cancer is largely dependent on common diagnostic procedures such as chest X-rays, sputum sample cytological investigations, optical fiber examinations of the bronchial airways, and complex procedures like CT and MRI scans. The study does, however, highlight the ongoing difficulties brought on by low specificity and sensitivity in CT and chest radiography.

Nithila and Kumar [5] present a novel active contouring model for lung segmentation, employing a variation level set function to enhance accuracy. Accurate diagnosis of lung diseases depends on precise segmentation of the lung parenchyma, and this study uses computed tomography (CT) imaging for image analysis. Significant progress has been made in CT lung image segmentation with the authors' introduction of the Gaussian filtering-new signed pressure force, a selective binary, and the SBGF-new SPF function. This novel method successfully detects the boundaries of the external lung and stops inefficient expansion at the margins. The suggested strategy's computational effectiveness and dependability are confirmed through comparative analyses with four different active contour models, establishing it as a viable approach for lung segmentation and disease detection.

Sangamithraa and Govindaraju [6] use a preprocessing strategy that includes median and Wiener filters to remove unwanted artifacts, improving data quality for subsequent analysis. The study uses the EK-mean clustering technique for clustering and the K-means method for CT image segmentation. Fuzzy EK-mean segmentation is used to extract significant properties from the images, such as contrast, homogeneity, area, correlation, and entropy. The effectiveness of the suggested methodology for lung cancer detection is demonstrated using a backpropagation neural network to complete the classification task [7].

Lanjewar *et al.,* [34] developed a novel methodology for detecting lung cancer by modifying the DenseNet201 model and using feature selection methods. Four different categories of lung cancer—adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal cell—were identified based on their study, which was conducted using the Kaggle chest CT-scan images dataset. There were improvements made to the DenseNet201 model, which included four more layers added to the original architecture. To extract the best features from the modified DenseNet201, two feature selection techniques were also used. These features were then applied to a variety of machine learning classifiers. Several evaluation metrics were used to fully evaluate the performance of their suggested system, including a confusion matrix, ROC curve, Kappa score (KS), 5-fold method, Cohen's Matthews Correlation Coefficient (MCC), and p-value. The system performed incredibly well, averaging 95% accuracy on average, and reaching 100% accuracy in certain instances. After a thorough 5-fold cross-validation, the p-value was found to be less than 0.001. The study conducted highlights the potential of their machine learning and computer technology approach to significantly improve the accuracy of lung cancer diagnoses made from CT scans.

In Table 4, a comparison of datasets utilized by different authors is provided, illustrating the variation in sources and contents of the datasets employed in diverse lung cancer detection investigations.

Table 4
Comparison of datasets

Reference	Dataset
Janee Alam et al. [30]	UCI ML Database
Maja Stella et al. [32]	ACDC, LUNGH
Gian Son Tran et al. [19]	LIDC-IDRI
Yutong Xie et al. [21]	LIDC-IDRI
M. B. Rodrigues et al. [31]	LIDC-IDRI
M. S. Rahman et al. [8]	TCIA
Rebecca L et al. [25]	Kaggle Data Science Bowl, LUNA 16
W. Chen et al. [29]	Shandong Cancer Hospital
Lanjewar et al. [34]	Kaggle chest CT-scan images dataset

A vast dataset that was obtained from Kaggle was used for this study. It included a variety of CT scan pictures that showed both normal instances and various forms of lung cancer. Large-cell carcinoma, squamous cell carcinoma, adenocarcinoma, and normal lung scans are all included in the carefully selected dataset. A comprehensive examination of the intricacies connected to specific lung cancer subtypes is ensured by the inclusion of these many forms of the disease. By exploiting this Kaggle dataset, the project intends to give useful insights into the intricate categorization and accurate detection of diverse kinds of lung cancer, pushing the boundaries of precision medicine in the field of lung cancer diagnostics.

2. METHODOLOGY

The entire research process can be divided into distinct phases in this study, which used a comprehensive methodology to advance the understanding and classification of lung cancer by using image processing in combination with machine learning techniques. Figure 6 displays the proposed approach's diagram.

2.1 Dataset Selection

A comprehensively curated dataset representing a range of lung cancer types, such as squamous cell carcinoma, large-cell carcinoma, adenocarcinoma, and normal cases, was assembled. Based on this dataset, a detailed analysis of the unique traits of every type of lung cancer will be conducted.

2.1.1 Dataset details

For this study, we have carefully chosen an extensive dataset that includes normal cases as well as squamous cell carcinoma, large-cell carcinoma, and adenocarcinoma, among other types of lung cancer. The dataset, which was obtained from Kaggle, includes 215 normal cases, 187 large-cell carcinoma images, 338 adenocarcinoma images, and 260 squamous cell carcinoma images. The diverse representation of the dataset served as the driving force behind this decision, enabling a detailed analysis of the unique characteristics linked to each subtype of lung cancer. Diversity is essential for developing strong classification models that can distinguish between various types of cancer with accuracy. Kaggle, a well-known source of high-quality datasets, guaranteed the data's dependability and accessibility.

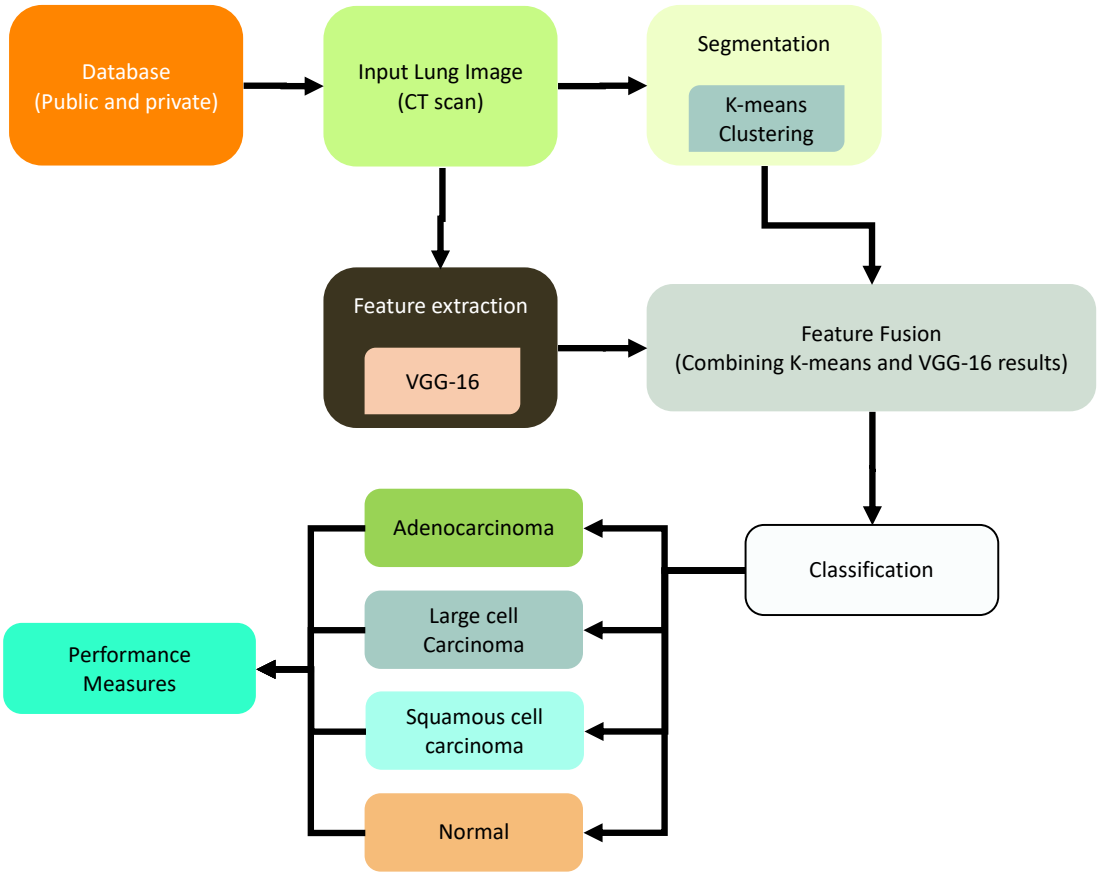


Fig. 6. Proposed approach

2.2 Pre-processing and Segmentation

Medical image processing involves the use of the segmentation method. A picture's fundamental function is to distinguish between elements that are advantageous and those that are detrimental. Because of this, it divides an image into discrete parts according to how much each part resembles the parts around it.

Utilizing K-means clustering, the dataset's images were successfully separated into several groups for segmentation [24]. The separation of pertinent features that were essential for further investigation was made possible by this segmentation.

This method is most frequently employed for medical image segmentation. The image is separated into several distinct groups, or clusters, that do not overlap with one another during the clustering process. There is no connection of any kind between these clusters. Using k reference points as a basis, the K-means clustering algorithm partitions the available data into k distinct groups.

2.3 Feature Extraction

The principal technique for extracting features was predicated on the VGG-16 framework. The VGG-16 model, which is shown in detail in Figure 7, underwent a rigorous optimization process to extract features related to lung cancer. By employing VGG-16's deep learning capabilities, the model was able to obtain intricate patterns associated with different types

of lung cancer. The VGG-16 model's architecture is visually represented in Figure 7, highlighting its complex layers and operations intended for efficient feature extraction in the context of lung cancer diagnosis.

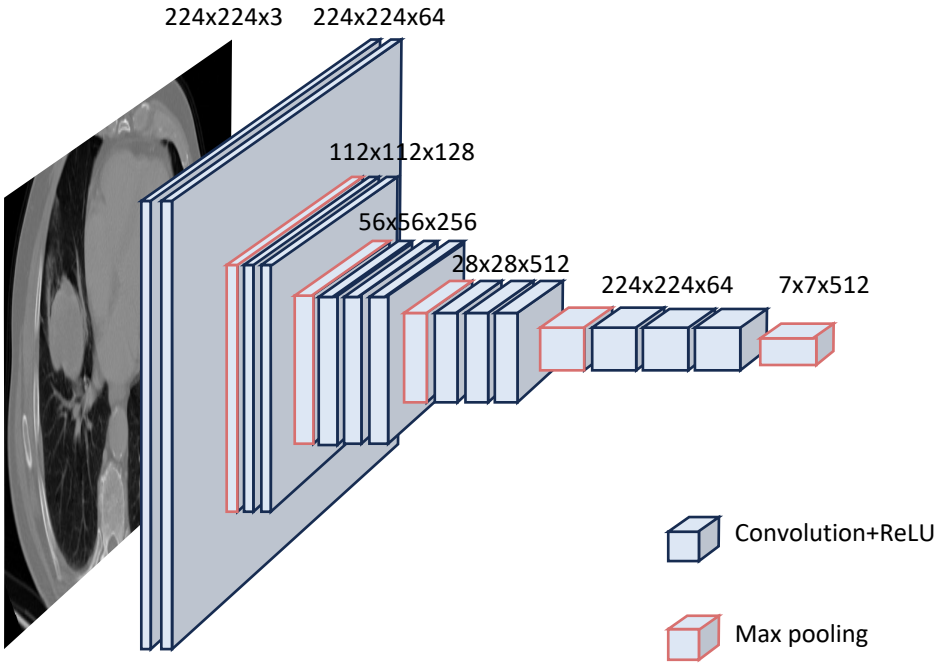
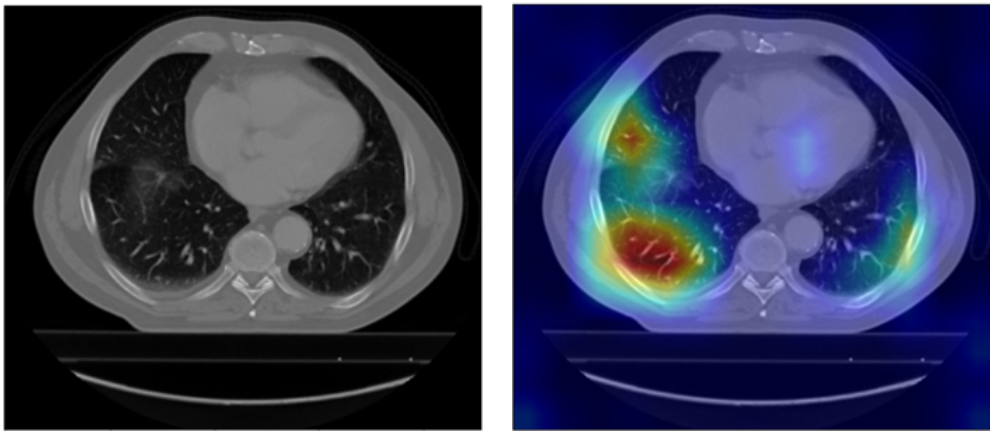


Fig. 7. Feature extraction architecture of VGG-16

The paradigm shift towards automated image interpretation is in line with the use of the VGG-16 architecture for feature extraction. VGG-16 is an image classification system that automatically recognizes pertinent patterns to mimic the diagnostic skills of medical professionals. This improves the interpretive capabilities of the model by reflecting the human ability to identify minute details characteristic of particular types of lung cancer.

2.3.1 Interpretability through Grad-CAM

To better understand why VGG-16 excels at feature extraction, we used Grad-CAM (Gradient-weighted Class Activation Mapping) for model interpretation. Grad-CAM produces heatmaps that show the areas of the input image that have a major influence on the model's decision-making. This method helps explain why VGG-16 concentrates on regions that are essential for precise lung cancer classification [38].



(a) (b)

Fig. 8. Feature extraction architecture of VGG-16 (a) Original CT scan image for lung cancer detection. (b) Grad-CAM heatmap highlighting regions crucial for VGG-16's decision-making.

The Grad-CAM heatmap draws attention to areas of the picture that VGG-16 deems essential for precise classification. We can interpret the complex patterns and structures that VGG-16 picks up with the aid of this interpretability tool, which supports our claim that VGG-16 is highly effective at extracting relevant features related to lung cancer. This visualization offers important insights into the feature extraction capabilities of VGG-16 and helps to explain why it focuses on specific regions.

Using Grad-CAM, we decipher the complex structures and patterns that VGG-16 has learned, highlighting the areas that are essential for precise classification. Our assertion that VGG-16 is skilled at extracting the best features related to lung cancer is supported by this interpretability tool.

2.4 Combination of Segmented and Extracted Features

A fused dataset was produced by combining the segmented images with the features that VGG-16 had extracted, preserving both spatial and deep learning-based information.

Using segmented images in conjunction with VGG-16 features extracted offers a comprehensive diagnostic method. This integration is the result of combining traditional radiological interpretations with advanced computational methods. By combining segmentation's spatial data with VGG-16's abstract features, the model improves diagnostic accuracy by developing a more sophisticated understanding of lung pathology.

2.5 Classification

The process of classification consists of two steps: building a model, which involves describing a set of predefined classes, and using the model to predict future or unknown objects [39]. Among the machine learning models used were logistic regression, support vector machines (SVM), and feedforward neural networks (FNNs). Using the segmented and feature-extracted dataset, each model was trained to discover the complex patterns linked to various forms of lung cancer.

These models were customized on an individual basis to reveal distinct aspects of the classification of lung cancer, thereby advancing a thorough and intricate understanding.

SVM, which is well known for its adaptability, functioned as a reliable classifier in the medical imaging dataset, showing exceptional skill in identifying complex patterns and nonlinear relationships [26]. Its ability to navigate high-dimensional medical data improved the ability to distinguish between different cancer subtypes, elevating the diagnostic process and assisting medical practitioners in accurately diagnosing diseases.

Logistic regression's ease of use and interpretability were utilized to identify correlations between the distinct forms of lung cancer in the medical images and the extracted features [27]. Logistic regression, a mainstay of statistical modeling, provided insightful information that helped medical professionals understand the nuances of classification results.

The FNN, which mimics human thought processes, has become a smart aide in the diagnostic field of medical imaging. The FNN's architecture was specifically designed to recognize complex patterns and variations that span the spectrum of lung cancer in medical images. This allowed it to translate visual cues into actionable insights, simulating a radiologist's cognitive approach and assisting in medical decision-making.

The Feedforward Neural Network (FNN) is a key player in the field of medical image classification [41] , especially when it comes to evaluating CT scans for lung cancer. The FNN is an effective tool for identifying complex patterns in medical images because of its layered architecture, which has three layers: an input layer, a hidden layer, an output layer. Compared to traditional ANNs, the architecture of the FNN can be customized to fulfil the unique requirements of medical diagnostics [28].

The hyperparameters of the SVM, logistic regression, and FNN models are presented in Table 5.

Table 5
Hyperparameters of the model

Model	Hyperparameter	Value
SVM	Kernel	Linear
	C	1.0
Logistic Regression	Solver function	ReLu
	Random State	42
FNN	Input Shape	input_shape
	Number of Hidden Units	128
	Activation Function	Relu
	Regularization	L2 (0.0001)
	Batch Normalization	Yes
	Dropout Rate	0.3
	Output Units	4
	Output Activation	Softmax
	Optimizer	Adam
	Loss	Sparse Categorical Crossentropy
	Metric	Accuracy
	Epochs	10
	Batch Size	32

One single hidden layer in the FNN architecture functions similarly to a targeted diagnostic analysis, effectively extracting critical features to differentiate between various forms of lung cancer seen in CT scans. On the other hand, adding more hidden layers improves the network's ability to recognize hierarchical representations, which makes it possible to distinguish minute differences between different cancer subtypes. Since a FNN's hidden layers are absent, direct feature mapping—which emphasizes speed and simplicity—is prioritized. This is important for obtaining real-time diagnostic insights in medical imaging applications. The meticulous choice of FNN architecture satisfies the complex requirements of deciphering complex patterns in CT scans, guaranteeing precise and effective medical diagnosis.

3. Result Analysis

In the experimental study, a dataset of CT images from patients with various cancer types and normal cases was used. Images are pre-processed and segmented using the K-means technique. Using this segmentation, the region of interest can be located. Both the segmented and extracted features were then applied, followed by feature extraction using a pre-trained model VGG16.The neural network and conventional models have both been trained using the transfer learning technique. Techniques for machine learning classification are then used.

3.1 Performance Measures

The models are evaluated using the following characteristics: F1-score, recall, accuracy, and precision. Together, these measures offer a thorough evaluation of how well the models perform in accurately categorizing various forms of lung cancer and differentiating between cases that are malignant and those that are not [37].

3.1.1 Accuracy

It measures how well a value corresponds to the available data. It is evaluated by using Eq. (1) [42].

$$Ae = \frac{TrN + TrP}{TrN + TrP + FaN + FaP} \quad (1)$$

where TrN for true negative, Trp for true positive, FaN for false negative and FaP stands for false positive.

3.1.2 Recall

It quantifies the percentage of real positive cases that the model accurately detected. by using Eq. (2) [42].

$$Recall = \frac{TrP}{TrP + FaN} \quad (2)$$

3.1.3 Precision

It quantifies the percentage of anticipated positive cases that the model accurately recognized. by using Eq. (3) [42].

$$Precision = \frac{TrP}{TrP+FaP} \tag{3}$$

3.1.4 F1-Score

The harmonic mean of recall and accuracy, which uses Eq. (4) to provide a balanced metric that takes into account both false positives and false negatives.

$$F1 - Score = \frac{2.Precision.Recall}{Precision+Recall} \tag{4}$$

Figures 9–11 present the outcomes of various machine learning predictors.

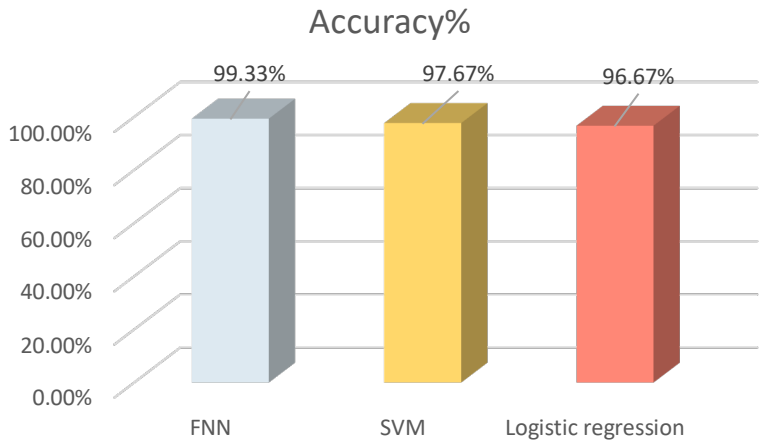


Fig. 9. Accuracy of machine learning methods for identifying lung cancer

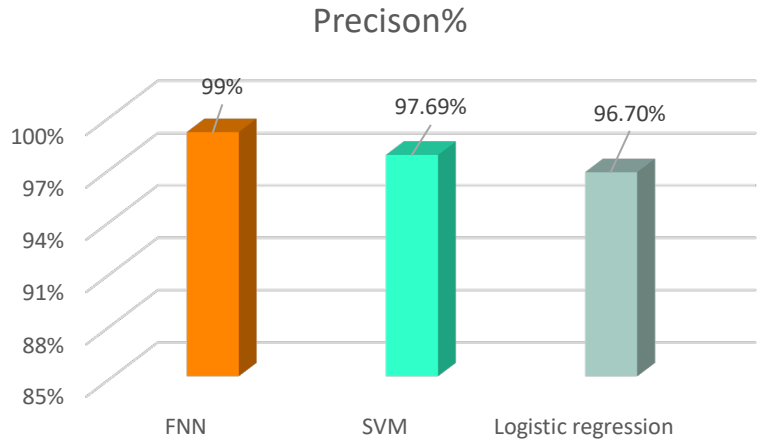


Fig. 10. Precision of machine learning methods for identifying lung cancer

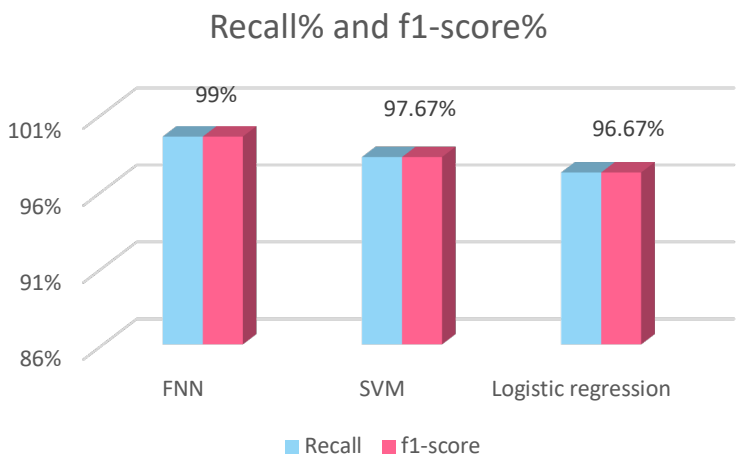


Fig. 11. Recall and f1-score of machine learning methods for the identification of lung cancer

3.2 K-fold Cross-Validation Results

To ensure the robustness and generalizability of our proposed models, we used a 5-fold cross-validation. To do this, we divided our dataset into five subsets. For each iteration, we trained the models on four of the subsets and validated on the fifth. After calculating the average performance scores, the procedure was repeated five times.

A balance between computational efficiency and accurate model performance estimation is obtained when selecting k=5 for cross-validation. Lowering the value of k, to say 5, guarantees a more robust evaluation at the same time as improving computational efficiency. Moreover, k=5 is a widely acknowledged selection in the literature that provides a reasonable trade-off between computational cost and variance reduction [33].

The average performance scores for the Feedforward Neural Network (FNN), Support Vector Machine (SVM), and Logistic Regression models obtained through 5-fold cross-validation are shown in Figure 12. Accuracy, Precision, Recall, and F1 Score are among the metrics.

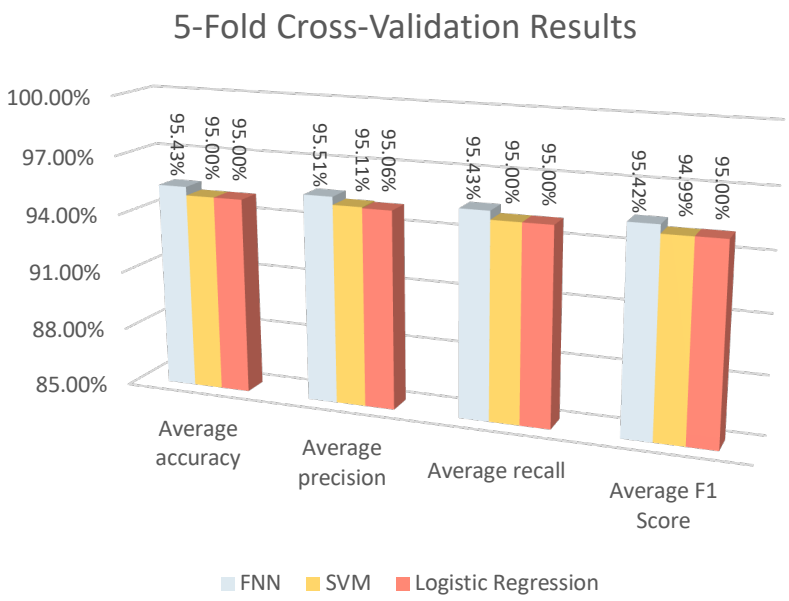


Fig. 12. Illustrates the average performance scores obtained from 5-fold cross-validation for Feedforward Neural Network (FNN), Support Vector Machine (SVM), and Logistic Regression models. The metrics include Accuracy, Precision, Recall, and F1 Score.

3.2 Overall result

The accuracy of Feedforward Neural Network (FNN) is better following the SVM and Logistic regression. FNN has performed better in every aspect having a great precision, recall and f1-score scoring 99.00%. Notably in 5-fold cross validation FNN has performed with better average accuracy of 95.43% including precision, recall and f1 score of 95.51%, 95.43% and 95.32% respectively. Table 6 clearly shows the performance measures of FNN, SVM and Logistic regression.

Table 6
Benchmarks for performance

Model	Precision	Recall	Accuracy	F1-score	5-fold cross-validation			
					Precision	Recall	Accuracy	F1-score
FNN	99%	99%	99.33%	99%	95.51%	95.43%	95.43%	95.42%
SVM	97.69%	97.67%	97.67%	97.67%	95.11%	95.00%	95.00%	94.99%
Logistic regression	96.70%	96.67%	96.67%	96.67%	95.06%	95.00%	95.00%	95.00%

Table 7 presents a thorough comparison of different models, including the suggested approach. For each model, the table displays the important metrics for performance, such as F1 score, recall, accuracy, and other attributes. With an accuracy of 99.33% and perfect scores of 99 in recall, precision, and F1 score, the suggested model stands out. This illustrates the higher effectiveness of our approach in accurately classifying lung cancer kinds. The suggested design is a levying contender to improve the field of lung cancer diagnostics since it can outperform alternative approaches, as the table illustrates.

Table 7
Comparing the approach and outcomes

Reference	Approach	Outcomes
W. Chen et al. [29]	CNN that is both 2D and 3D with a hybrid fusion module (HFFM)	Dice score - 88.8, Sensitivity - 87.2, Precision - 90.9
M. B. Rodrigues, et al. [31]	Laplace, Gaussian & Sobel filtering, SVM, KNN, SCM Mean HU, multilayer perceptrons	SCM Mean HU Accuracy: 96.70
Yutong Xie et al. [21]	Collaborative Deep Learning based on knowledge, U-Net, and 3D-GLCM-SVM	Accuracy - 91.60%, Specificity - 94%, AUC - 95.70%, Sensitivity - 86.52%
Gian Son Tran et al. [19]	2D Deep Convolutional Network	Accuracy - 97.2, Sensitivity - 96.0, Specificity - 97.3
Rebecca L et al. [25]	3D Probabilistic Deep Learning, V-Net architecture	CADe sensitivity - 96.5%, CADx AUC - 0.87
Maja Stella et al. [32]	VGG16, ResNet50, CNN	Accuracy - 97.9, 93
Janee Alam et al. [30]	Watershed Transform, GLCM	Identification - 97, Cancer Prediction - 87
M. S. Rahman et al. [8]	Gaussian Blur, Otsu Threshold, Inception-V3, VGG-8, and MobileNet	Accuracy - 97%, Sensitivity - 96.26%,Specificity - 97.85%
Lanjewar et al. [34]	Modified DenseNet201, Feature selection methods	Average accuracy - 95
V. Nisha Jenipher et al. [35]	R-CNN with MobileNetV2 and SCAM framework	Accuracy - 98.6%, Specificity - 96.8%, Sensitivity - 97.5%, Precision - 98.2%
Vani Rajasekar et al. [36]	VGG16,Resnet50,InceptionV3	Accuracy-96.52,93.47,93.54 Precision-92.14,93.31,90.57
Proposed Approach	Kmeans, VGG16 feature extraction, Fusion,SVM,Logistic Regression,FNN	FNN Accuracy-99.33,Precision-99,Recall-99,F1-score-99 Cross validation average- accuracy-95.43,precision-95.51,recall-95,F1-score-95.42

4. CONCLUSION

Lung cancer remains a major threat to people's lives all over the world and a major challenge to global health. Because of its high death rate, prompt detection is essential for successful treatment. Lung cancer's complexity necessitates novel approaches, and in this regard, combining image processing and machine learning presents a state-of-the-art remedy. Using cutting-edge methods, this study explores the complex classification of lung cancer types, aiming to differentiate between cancerous and non-cancerous cases as well as between particular cancer subtypes.

With today's complicated medical technologies, a precise and comprehensive diagnosis of lung cancer is more important than ever. Promising outcomes have been observed in the utilization of machine learning models, including the Feedforward Neural Network (FNN). The FNN is a strong tool for classifying lung cancer because of its impressive accuracy of 99.33%. The model's ability to attain remarkable recall, f1-score, precision and high accuracy places it at the forefront of the search for cutting-edge diagnostic techniques.

The role that technology plays in enhancing diagnostic capabilities is becoming more and more apparent as we navigate the difficulties presented by lung cancer. The remarkable performance of the FNN highlights the possibility of a paradigm change during the lung cancer

diagnosis process. By advancing medical technology, this study contributes to the continuing conversation about improving lung cancer early detection, categorization, and intervention, offering hope for better patient outcomes and fewer fatalities.

Acknowledgement
This research was not funded by any grant.

References

[1] Palani, D., and K. Venkatalakshmi. 2018. "An IoT Based Predictive Modelling for Predicting Lung Cancer Using Fuzzy Cluster Based Segmentation and Classification." *Journal of Medical Systems* 43 (2): 21. https://doi.org/10.1007/s10916-018-1139-7

[2] Nageswaran, Sharmila, G. Arunkumar, Anil Kumar Bisht, Shivlal Mewada, J. N. V. R. Swarup Kumar, Malik Jawarneh, and Evans Asenso. 2022. "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing." *BioMed Research International* 2022: 1755460. https://doi.org/10.1155/2022/1755460

[3] Geng, Lei, Siqi Zhang, Jun Tong, and Zhitao Xiao. 2019. "Lung Segmentation Method with Dilated Convolution Based on VGG-16 Network." *Computer Assisted Surgery (Abingdon, England)* 24 (sup2): 27–33. https://doi.org/10.1080/24699322.2019.1649071

[4] Saini, Ashwini Kumar, H. S. Bhadauria, and Annapurna Singh. 2016. "A Survey of Noise Removal Methodologies for Lung Cancer Diagnosis." In *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. IEEE. https://doi.org/10.1109/CICT.2016.139

[5] Joon, P., S. B. Bajaj, and A. Jatain. 2019. "Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Tech- Niques." In *Progress in Advanced Computing and Intelligent Engineering*, 13–23. Singapore: Springer. https://doi.org/10.1007/978-981-13-1708-8_2

[6] Sangamithraa, P. B., and S. Govindaraju. 2016. "Lung Tumour Detec- Tion and Classification Using EK-Mean Clustering." In *Proceed- Ings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking*. Chennai, India. https://doi.org/10.1109/WiSPNET.2016.7566533

[7] Kurkure, Manasee, and Anuradha Thakare. 2016. "Lung Cancer Detection Using Genetic Approach." In *2016 International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE. https://doi.org/10.1109/ICCUBEA.2016.7860007

[8] Rahman, M. S., P. C. Shill, and Z. Homayra. 2019. "A New Method for Lung Nodule Detection Using Deep Neural Networks for CT Images Int." In *Conf. on Electrical*, 1–6. https://doi.org/10.1109/ECACE.2019.8679439

[9] Deepa, N., B. Prabadevi, Praveen Kumar Maddikunta, Thippa Reddy Gadekallu, Thar Baker, M. Ajmal Khan, and Usman Tariq. 2021. "An AI-Based Intelligent System for Healthcare Analysis Using Ridge-Adaline Stochastic Gradient Descent Classifier." *The Journal of Supercomputing* 77 (2): 1998–2017. https://doi.org/10.1007/s11227-020-03347-2

[10] Qu, G., D. Zhang, and P. Yan. 2001. "Medical Image Fusion by Wavelet Transform Modulus Maxima." *Optics Express* 9 (4): 184–90. https://doi.org/10.1364/OE.9.000184

[11] Tekade, R. 2018. "Lung Nodule Detection and Classification Using Machine Learning Techniques." *Asian Journal for Convergence in Technology* 4. https://doi.org/10.1109/ICCUBEA.2018.8697352

[12] Sone, S., S. Takashima, F. Li, Z. Yang, T. Honda, Y. Maruyama, M. Hasegawa, et al. 1998. "Mass Screening for Lung Cancer with Mobile Spiral Computed Tomography Scanner." *Lancet* 351 (9111): 1242–45. https://doi.org/10.1016/S0140-6736(97)08229-9

[13] Kuruvilla, Jinsa, and K. Gunavathi. 2014. "Lung Cancer Classification Using Neural Networks for CT Images." *Computer Methods and Programs in Biomedicine* 113 (1): 202–9. https://doi.org/10.1016/j.cmpb.2013.10.011

[14] Lee, Michael C., Lilla Boroczky, Kivilcim Sungur-Stasik, Aaron D. Cann, Alain C. Borczuk, Steven M. Kawut, and Charles A. Powell. 2010. "Computer-Aided Diagnosis of Pulmonary Nodules Using a Two-Step Approach for Feature Selection and Classifier Ensemble Construction." *Artificial Intelligence in Medicine* 50 (1): 43–53. https://doi.org/10.1016/j.artmed.2010.04.011

[15] Orozco, H. M., O. V. Villegas, G. C. Sánchez, J. O. Domínguez, and J. N. Alfaro. 2014. "Automated System for Lung Nodules Classification Based on Wavelet Feature Descriptor and Support Vector Machine." *Biomedical Engineering Online* 14 (1). https://doi.org/10.1186/s12938-015-0003-y

[16] Chan, Heang-Ping, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. 2020. "Deep Learning in Medical Image Analysis." *Advances in Experimental Medicine and Biology*1213: 3–21. https://doi.org/10.1007/978-3-030-33128-3_1

[17] Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. 2018. "Artificial Intelligence in Radiology." *Nature Reviews. Cancer* 18 (8): 500–510. https://doi.org/10.1038/s41568-018-0016-5

[18] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/CVPR.2017.369

[19] Tran, Giang Son, Thi Phuong Nghiem, Van Thi Nguyen, Chi Mai Luong, and Jean-Christophe Burie. 2019. "Improving Accuracy of Lung Nodule Classification Using Deep Learning with Focal Loss." *Journal of Healthcare Engineering* 2019: 5156416. https://doi.org/10.1155/2019/5156416

[20] Hirsch, Fred R., Giorgio V. Scagliotti, James L. Mulshine, Regina Kwon, Walter J. Curran Jr, Yi-Long Wu, and Luis Paz-Ares. 2017. "Lung Cancer: Current Therapies and New Targeted Treatments." *Lancet* 389 (10066): 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8

[21] Xie, Yutong, Yong Xia, Jianpeng Zhang, Yang Song, Dagan Feng, Michael Fulham, and Weidong Cai. 2019. "Knowledge-Based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT." *IEEE Transactions on Medical Imaging* 38 (4): 991–1004. https://doi.org/10.1109/TMI.2018.2876510

[22] Ginneken, Bram van, Cornelia M. Schaefer-Prokop, and Mathias Prokop. 2011. "Computer-Aided Diagnosis: How to Move from the Laboratory to the Clinic." *Radiology* 261 (3): 719–32. https://doi.org/10.1148/radiol.11091710

[23] Li, Guanzhen, Zongxia Li, Wenxu Shuai, and Yue Wang. 2022. "Optimizing VGG16 for Lung Cancer CT Image Recognition: Evaluating the Effectiveness of Channel and Spatial Attention." In *International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021)*, edited by Ke Chen, Nan Lin, Romeo Meštrović, Teresa A. Oliveira, Fengjie Cen, and Hong-Ming Yin. SPIE. https://doi.org/10.1117/12.2628043

[24] Pham, D. L., C. Xu, and J. L. Prince. 2000. "Current Methods in Medical Image Segmentation." *Annual Review of Biomedical Engineering* 2 (1): 315–37. https://doi.org/10.1146/annurev.bioeng.2.1.315

[25] Ozdemir, Onur, Rebecca L. Russell, and Andrew A. Berlin. 2019. "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans." *arXiv [Cs.CV]*. https://doi.org/10.1109/TMI.2019.2947595

[26] Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning*20 (3): 273–97. https://doi.org/10.1007/BF00994018

[27] Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Nashville, TN: John Wiley & Sons. https://doi.org/10.1002/9781118548387

[28] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539

[29] Chen, Wei, Haifeng Wei, Suting Peng, Jiawei Sun, Xu Qiao, and Boqiang Liu. 2019. "HSN: Hybrid Segmentation Network for Small Cell Lung Cancer Segmentation." *IEEE Access: Practical Innovations, Open Solutions* 7: 75591–603. https://doi.org/10.1109/ACCESS.2019.2921434

[30] Alam, J., and S. Hossan. 2018. "Multi-Stage Lung Cancer Detection and Prediction Using Multi-Class Svm Classifier Int." *Conf. on Computer, Communication, Chemical, Material and Electronic Engineering*, no. IC4ME2: 1–4. https://doi.org/10.1109/IC4ME2.2018.8465593

[31] Rodrigues, M. B., R. V. Da Nóbrega, S. S. Alves, Rebouças Filho, P. P. Duarte, J. B. Sangaiah, and De Albuquerque. 2018. "Health of Things Algorithms for Malignancy Level Classification of Lung Nodules IEEE." *IEEE Access* 6: 18592–601. https://doi.org/10.1109/ACCESS.2018.2817614

[32] Šarić, M., M. Russo, and Stella M. Sikora. 2019. "CNN-Based Method for Lung Cancer Detection in Whole Slide Histopathology Images Int." In *Conf. on Smart and Sustainable Technologies (SpliTech)*, 1–4. https://doi.org/10.23919/SpliTech.2019.8783041

[33] Lanjewar, Madhusudan G., Kamini G. Panchbhai, and Lalchand B. Patle. 2024. "Fusion of Transfer Learning Models with LSTM for Detection of Breast Cancer Using Ultrasound Images." *Computers in Biology and Medicine* 169 (107914): 107914. https://doi.org/10.1016/j.compbiomed.2023.107914

[34] Lanjewar, Madhusudan G., Kamini G. Panchbhai, and Panem Charanarur. 2023. "Lung Cancer Detection from CT Scans Using Modified DenseNet with Feature Selection Methods and ML Classifiers." *Expert Systems with Applications* 224 (119961): 119961. https://doi.org/10.1016/j.eswa.2023.119961

[35] Jenipher, V. Nisha, and S. Radhika. 2024. "Lung Tumor Cell Classification with Lightweight mobileNetV2 and Attention-Based SCAM Enhanced Faster R-CNN." *Evolving Systems*. https://doi.org/10.1007/s12530-023-09564-3
[36] Rajasekar, Vani, M. P. Vaishnnave, S. Premkumar, Velliangiri Sarveshwaran, and V. Rangaraaj. 2023. "Lung Cancer Disease Prediction with CT Scan and Histopathological Images Feature Analysis Using Deep Learning Techniques." *Results in Engineering* 18 (101111): 101111. https://doi.org/10.1016/j.rineng.2023.101111
[37] Deepapriya, B. S., Parasuraman Kumar, G. Nandakumar, S. Gnanavel, R. Padmanaban, Anbarasa Kumar Anbarasan, and K. Meena. 2023. "Performance Evaluation of Deep Learning Techniques for Lung Cancer Prediction." *Soft Computing* 27 (13): 9191–98. https://doi.org/10.1007/s00500-023-08313-7
[38] Lanjewar, Madhusudan G., Kamini G. Panchbhai, and Panem Charanarur. 2024. "Small Size CNN-Based COVID-19 Disease Prediction System Using CT Scan Images on PaaS Cloud." *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-17884-4
[39] Esmail, Fahd Sabry, Mohamed Badr Senousey, and Mohamed Ragaie Sayed. 2020. "Selecting the Best Model Predicting Based Data Mining Classification Algorithms for Leukemia Disease Infection". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 7 (1):1-10. https://akademiabaru.com/submit/index.php/araset/article/view/1927.
[40] Jamlos, Mohd Aminudin and Wan Azani Mustafa. "Improved Confocal Microwave Imaging Algorithm for Tumor Detection." (2019)
[41] Parveen, Rahila, Mairaj Nabi, Fayaz Ahmed Memon, Sabina Zaman and M Ali. "A Review and Survey of Artificial Neural Network in Medical Science." (2016)
[42] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011)