International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)
Vol. 11-No.3 (2025) pp. 3871-3880
http://www.ijcesen.com
ISSN: 2149-9144

Research Article

Optimized AI-Based Detection of Pulmonary Nodules Using VGG16 and XGBoost

Mithun B Patil[1]*, Tan Kuan Tak[2], Pravin R. Kshirsagar[3], R Thiagarajan[4], Sivaneasan Bala Krishnan[5]

[1]Associate Professor, N K Orchid College of Engineering and Technology, Solapur Maharashtra, India
* **Corresponding Author Email:** drmithunbpatil@gmail.com - **ORCID:** 0000-0001-7741-3645

[2]Associate Professor Engineering Cluster, Singapore Institute of Technology, Singapore
Email: KuanTak.Tan@SingaporeTech.edu.sg - **ORCID:** 0000-0002-0578-7098

[3]Professor & Head-ETC, J D College of Engineering &Management, Nagpur, India,
Email: pravinrk88@yahoo.com - **ORCID:** 0000-0002-7381-6284

[4]Professor, Department of Information Technology, VelTech MultiTech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai – 62
Email: rthiyagarajantpt@gmail.com - **ORCID:** 0000-0003-2107-5047

[5]Associate Professor Engineering Cluster, Singapore Institute of Technology, Singapore
Email: sivaneasan@SingaporeTech.edu.sg - **ORCID:** 0000-0002-4271-2677

Article Info:

DOI: 10.22399/ijcesen.2122
Received : 04 March 2025
Accepted : 07 May 2025

Keywords :

Lung Cancer Detection
Ensemble Learning
VGG16, Histogram of Oriented Gradients (HOG)
XGBoost, Transfer Learning
Convolutional Neural Networks (CNN)
Image Augmentation

Abstract:

A novel approach for earlier lung cancer detection is combining the XGBoost classifier with a Histogram Focused on Gradients (HOG) data and VGG16, a deep learning model. With a startling accuracy range of 97–98%, our ensemble-based approach clearly beats conventional techniques. Celebrated for its ability to catch minute details in images, the VGG16 model stresses structural components and performs quite elegantly with HOG properties. Feeding these data into the XGBoost classifier—known for speed and performance—results in a quite accurate and dependable lung cancer detection tool. Our results clearly show the benefits over traditional approaches; thus, they emphasize the likely usage of this new approach in clinical environments. Given its high accuracy, our model seems to be a useful instrument for early lung cancer detection, which is essential to raising patient survival rates and outcomes. Future larger datasets will allow us to test our model more thoroughly and investigate its integration into clinical procedures. Moreover, evaluating the model on other imaging modalities and different patient groups will help to guarantee its general relevance and strength. In the end, this research provides the means for the construction of sophisticated diagnostic instruments able to change early lung cancer diagnosis and, therefore, improve the general results of treatment.

1. Introduction

Lung cancer is among the most frequent and deadly cancers worldwide, and it remains a major and significant public health issue. It is one of the biggest tasks to diagnose lung cancer in its primary stages and to classify the various types of cancer accurately even today, when medical technologies and treatments are far more advanced. Timing is critical in determining the success of treatment, and this is why early diagnosis plays a substantial role in the enhancement of the well-being of the patient. However, it remains a great clinical problem to delineate the various types of lung lesions, such as benign, malignant and normal ones. The need for novel diagnostic strategies capable of accurately distinguishing between these groups remains perhaps even more significant than it was before. Progress in the field of ML and AI makes it possible to create new diagnostic support tools that can analyse medical imaging data faster and more accurately than a human can [2]. In the field of medical image investigation, deep learning and, more specifically, CNN has been shown to be some of the most effective AI tools. CNNs are very effective in capturing complex patterns from the data, especially for problems like image classification and segmentation. Among the most promising approaches in deep learning is the transfer learning, a method in which the network is fine-tuned for a certain dataset. VGG16 is perhaps the most popular deep learning model that was pre-trained on the large-scale ImageNet dataset and has been known to work well on a range of medical imaging challenges. This make it particularly suitable for lung cancer classification due to its ability to extract and use high level features. For instance, but extracting the features of VGG16 and training it further for lung cancer diagnosis, researchers have recorded even higher accuracies. Other conventional machine learning methods that have also been applied in the diagnosis of lung cancer include the SVM, random forest, and XGBoost classifiers. These models are usually learned with the help of such methods as Histogram of Oriented Gradients (HOG) features, which contain important shape and texture information extracted from medical images.

While traditional classification methods may not necessarily yield accuracy as high as deep learning models, they have certain strengths such as interpretability and computational costs. However, there are ensemble methods that combine several classifiers to improve diagnostic accuracy have been developed. The bagging and boosting techniques combine the results of individual classifiers making them more accurate and stable. If the predictions of the models are combined, ensemble methods can help overcome individual classifiers' flaws and offer more accurate diagnostic results. The objectives of this work are as follows: The global goal of this research is to design a diagnostic framework that integrates deep learning and conventional machine learning algorithms to classify lung lesions into benign, malignant, and normal categories. The proposed method combines CNN with transfer learning technique with VGG16 network for feature extraction and classification alongside conventional method like SVM, RF and XGB trained on HOG features. The concluding diagnosis is made using an ensemble approach whereby the outcomes of the models are aggregated through majority voting. In order to test the proposed diagnostic framework, a dataset covering medical images of lung lesions is used. This way, we have a sensible set of benign, malignant, and normal cases, allowing to provide a systematic evaluation of the models. These images go through preprocessing and data augmentation to make the models more robust [11]. Next, the dataset is divided into training and testing data to ensure proper evaluation of the system. For evaluating the performance of the proposed method, accuracy, precision, and recall. To evaluate the effectiveness of the proposed method, a comparative study is performed with other classifiers and prevailing diagnostic techniques. This work turns into the current literature on lung cancer diagnosis, offering an integrated deep learning and machine learning model that can enhance the accuracy of detection. The results obtained in this study could be useful to improve the diagnostic techniques and to obtain better results in the early diagnosis of lung cancer patients.

One might regard this study as a significant addition to lung cancer diagnosis tool. When compared to existing techniques for the categorisation of lung lesion, the suggested diagnostic framework combines the deep learning approach with standard machine learning technique, which has favourable aspects. This method increases the stability and transferability of models in addition to lowering the misdiagnosis rate of lung cancer. Consequently, this study could have a very important relevance for clinical practice since it would help the doctors to make better decisions and hence improve the condition of the patients. All things considered, this paper emphasises the value and part artificial intelligence play in the medical field. Jiang and colleagues 2023 Along with the ways of patient treatment and general management of care, these technologies will progressively become more included into diagnosis and treatment of diseases as they develop [6]. Apart from supporting doctors and researchers in their attempts to improve methods of lung cancer diagnosis, this study offers the path for the future development of many additional approaches of using medical imagery for diagnosis.

2. Related Work

Research on lung cancer detection is still very important since many studies investigate several approaches to improve the accuracy and efficiency of identifying malignant pulmonary nodules. With an eye on the integration of deep learning and machine learning approaches, this literature review explores the major contributions made by scholars in this field.

Particularly in cases of sarcoma, Baidya Kayal et al. (2023) suggested a thorough technique using computed tomography (CT) imaging combined with artificial intelligence (AI) models to evaluate the malignant potential of lung nodulesv [1]. Their research emphasises the growing relevance of artificial intelligence in medical imaging by demonstrating how these models can greatly help in early diagnosis by seeing minor trends in imaging data that might be missed by conventional approaches. Analogously, Canayaz et al. (2024) investigated deep learning techniques for the segmentation and classification of pulmonary nodules in chest CT data [2]. Their analysis underlined the advantages and drawbacks of several deep learning architectures and underlined the important part precise segmentation plays in enhancing diagnostic results. This work also highlighted the difficulties in teaching these models, particularly in relation to complicated and high-dimensional medical imagery. Excellent in the segmentation, identification, and classification of lung nodules, Dutande et al. (2021) presented an original 2D-3D cascaded CNN method. More complex analysis of nodules—which can change greatly in size and appearance over many slices of CT scans—is made possible by the cascading of 2D and 3D networks [3]. This method was proved to efficiently combine multi-dimensional data, hence enhancing detection accuracy. Based on metabolic parameters, Guan et al. (2023) showed how well XGBoost, a strong ensemble learning method, predicted early lung cancer [4]. Their study shows how flexible XGBoost is in managing different kinds of biomedical data and implies that, in medical diagnostics, ensemble models can have better predictive performance than single machine learning models.

Emphasising the fast developments in this sector, Javed et al. (2024) carefully reviewed deep learning methods for lung cancer detection [5]. Their article offered a thorough summary of the several deep learning models applied in cancer diagnosis, including with a discussion of the difficulties of using these models in clinical practice, such the need of extensive annotated datasets and the interpretability of model predictions.

Jiang et al. (2023) concentrated on the application of these models to medical imaging, thereby investigating the function of deep learning in cancer detection [6]. With an eye towards model transparency and the integration of domain information to raise diagnostic accuracy, they explored several architectures and methods applied in the diagnosis of several forms of cancer.

Deep convolutional neural networks (CNNs) were first used by Li et al. (2016) to classify CT imaging pulmonary nodules. Their early work prepared the way for later research by proving CNNs' ability to automatically learn and extract features from medical images without requiring significant hand feature engineering [7]. The use of pre-trained networks and the bag-of- features technique for the automatic categorisation of lung nodules was investigated in the 2023 Lima et al. paper This work demonstrated how transfer learning may be applied in medical image analysis such that models might incorporate knowledge from other fields and raise their performance on somewhat modest medical datasets [8]. An attention-based deep learning network meant to distinguish benign from malignant lung nodules was presented by Liu et al. (2023) [9]. This method proved great accuracy in malignancy classification by concentrating the model's attention on the most pertinent areas of the image, therefore illustrating the advantages of attention processes in medical image analysis. Developing a 3D attention-gated convolutional network that considers related pulmonary fibrosis, a major consequence in lung cancer patients, Liu et al. (2024) expanded this line of study [10]. Their model added extra contextual information from 3D CT scans to raise classification accuracy even more. Niranjan Kumar et al. (2021) underlined in their analysis of the UNet architecture the need of precise lung nodule segmentation [11]. This work underlined how exact segmentation is essential for next stages in the diagnostic pipeline, such classification and analysis, and how it can directly affect the general performance of the detection system. 2020's Riquelme and Akhloufi thoroughly examined deep learning methods for CT scan lung cancer nodule identification and categorisation [12]. Their study addressed the development of deep learning models in this field as well as the several approaches applied to address the particular difficulties presented by medical picture data, like imbalance in class distribution and the necessity of extensive labelled datasets.

In 2023 Shah et al. presented an ensemble method using many 2D CNN models for lung cancer diagnosis [13]. Particularly in relation to diverse and challenging medical datasets, their study showed how ensemble approaches may increase model resilience and generalisation.

Shen et al.'s (2015) multi-scale CNN model likewise significantly advanced this field by gathering features at various layers inside the image [14]. This technique proved particularly effective in spotting nodules of various sizes, a common challenge in the detection of lung cancer. Shimazaki et al. (2022) developed a deep learning-based method using segmenting techniques and chest radiographs by means of which the lung sections could be detected before classification [15]. Their approach underscored the need of preprocessing phases in increasing the accuracy of AI-based diagnostic tools. Using fresh deep belief network (DBN) suggested by Siddiqui et al. gabor filters (2020), lung cancer detection and classification is recommended to be supported [16]. This work demonstrated how hybrid approaches by integrating modern deep learning models with classic image processing techniques may increase model performance. Investigating new deep learning techniques for lung cancer identification helps Wankhade and Vigneshwari's (2023) study to show ongoing improvement and the promise of artificial intelligence to transform cancer diagnosis and so strengthens the area [17]. Using transfer learning with CNNs and random forests for lung nodule classification, Yahya Saleh et al. (2023) demonstrated how combining deep learning with traditional machine learning approaches may produce higher results, particularly in circumstances with limited training data [18]. To solve class imbalance in pulmonary nodule classification, Yi et al. (2023) presented a multi-label softmax network considering category dependencies [19]. Their efforts help to define how effectively to handle the complexity of medical picture data. With an eye towards especially enhancing the identification of tiny lesions, both Zheng et al. (2021) and Zheng et al. (2020) helped to create deep CNNs for the diagnosis of lung cancer [20, 21]. Their studies showed how CNNs might raise early detection rates by concentrating on difficult scenarios where conventional approaches usually fall short.

Particularly by use of deep learning and machine learning approaches, this large body of research emphasises the notable developments achieved in the field of lung cancer diagnosis. Combining VGG16 for feature extraction with Histogram of Orientated Gradients (HOG) features with XGBoost, the proposed ensemble-based strategy seeks to improve on these developments by using the benefits of both deep learning and classical machine learning techniques. This method is supposed to increase lung cancer detection accuracy and resilience, hence improving clinical practice outcomes.

Several important difficulties still exist even with the great developments in deep learning and machine learning-based lung cancer detection. For example, even if feature extraction has advanced thanks to research, more efficient approaches to blend deep learning elements with classic handcrafted ones are still needed to increase accuracy. Additionally, even though model interpretability is crucial, especially in medical contexts, many existing models are still too complex for doctors to fully understand and trust. Another ongoing issue is the detection of small pulmonary nodules, which are critical for early cancer diagnosis but often go unnoticed by current models. Notwithstanding continuous attempts to solve the issues of class imbalance and limited data, they still impair the performance of many models. Moreover, most studies concentrate on particular datasets, which restricts the generalisability of their results to more general populations. Finally, even if many experimental models show promise, their actual relevance in clinical environments greatly differs from their study results.

In section 3 we have developed A Novel Ensemble-Based Approach for Lung Cancer Detection Using VGG16 and HOG Features with XGBoost to fill in these voids. This method seeks to efficiently combine deep learning with conventional machine learning methods thereby improving the accuracy, transparency, and dependability of lung cancer detection.

3. Methodological Approach

3.1 Preliminaries

Before we dive into the proposed diagnostic framework, let us begin by first understanding the fundamentals of some of the key techniques on which we will build a strong foundation for it. This section gives a brief introduction to the CNNs, traditional machine learning techniques and ensemble methods which build up our example.

3.1.1 Convolutional Neural Networks (CNNs)

Comprising convolutional and pooling layers followed by one or more fully connected layers to enable the network to learn features from different levels of abstraction, serialised CNNs are an architectural class of deep learning models highly efficient for classification problems with images.

3.1.2 Transfer Learning

Here you build on top of a pre-trained network on a vast dataset, say ImageNet, and fine-tune it to your own little dataset. This approach reduces the time and data needed to reach higher performance on the task of interest by using acquired characteristics from the pre-trained network.

3.1.3 Traditional Machine Learning Techniques

Support vectors machines (SVM) are supervised learning classification models. Usually, they discover the hyperplane in their feature space that divides every class the best.

Random forests are ensemble learning techniques whereby several decision trees are generated during training and output the mode of the classes for classification (Source). Tianqi Chen created XGBoost, Extreme Gradient Boosting, a scalable and accurate application of gradient boosting machines (GBMs). Most classification problems show XGBoost's better performance.

3.1.4 Histogram of Oriented Gradients (HOG) Features

HOG is a feature descriptor in computer vision and image processing applied for object detection and here it gets fascinating. This explains the strength and direction of the edge, so providing good shape and texture information.

3.1.5 Ensemble Methods

Ensemble techniques merge the forecasts of several models to raise general performance. A basic yet efficient ensemble technique where the final forecast is the one with most votes from individual models is majority voting.

3.2 Advanced Architectural Design for Lung Lesion Classification

For higher classification accuracy and resistance, the suggested method to classify lung lesion uses a novel diagnostic framework combining deep learning with conventional machine-learning approaches. Three main sections comprised this all-encompassing method: CNN-based feature description and classification; traditional machine learning models derived from HOG features and ensemble learning.

3.2.1 CNN-Based Feature Extraction and Classification

The VGG16 model, is pre-trained on the ImageNet dataset, drives the deep learning component at its core using a Convolutional Neural Network (CNN). This model is good in finding complex patterns in images because of its hierarchical structure including convolutional and pooling layers. These layers define high-level features from the lung lesioned input images. This work employs the VGG16 model to fine-tune its pre-trained features to the specific goal of lung lesion classification. Fine-tuning consists in removing the top layers of VGG16 and adding custom fully connected layers aimed to classify images into categories like benign, malignant, and normal categories. This method increases the capacity of the model to exploit pre-existing knowledge while changing to the new dataset, hence improving classification performance with limited computational resources and data.

3.2.2 Traditional Machine Learning Models with HOG Features

To deep learning algorithms thus rely HOG feature-based traditional machine learning methods. HOG is very good descriptor by which one can represent the essential shape and texture characteristics of the image and is particularly useful in the context of lung lesions distinguishing between different types of lung lesions. After extracting said features from the unilateral lesion images, the three classifiers of interest included the Support Vector Machines, Random Forests, and XGBoost. Each with its respective advantages: SVMs tend to separate classes by finding an optimal hyperplane, Random Forests provide strength in numbers by producing an ensemble of decision trees, and XGBoost is known to be very efficient and performing by applying gradient boosting to boost classification even more.

3.2.3 Ensemble Learning through Majority Voting

Integrating the predictions from the CNN-based model and the conventional classifiers, ensemble learning marks the last phase of the proposed approach. A majority vote system helps to attain this integration. Under this approach, every model votes for the anticipated class of a lung lesion; the majority vote determines the final classification. Combining the qualities of several models helps to reduce their shortcomings, therefore producing a more accurate and dependable diagnosis conclusion.

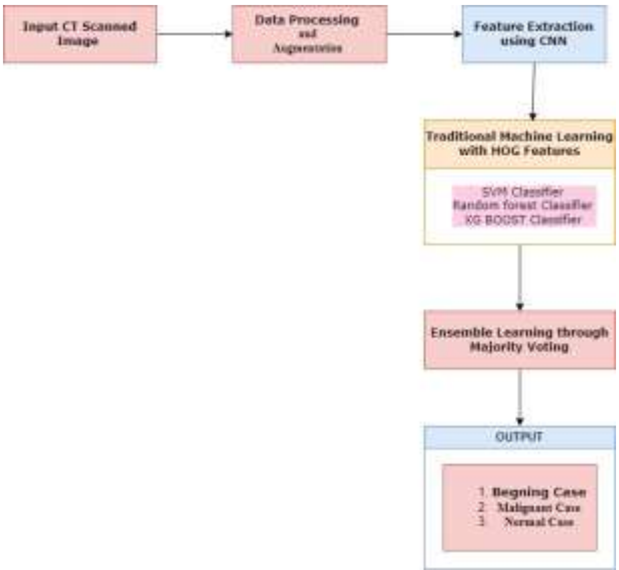

Figure 1. *Architecture of Proposed Method*

Step 1: Data Preprocessing and Augmentation

First, it requires an unbiased distribution of photographs of lung symptoms between benign, malignant, and normal cases. Preprocessing steps involve resizing images to some common dimensions (e.g., 224 by 224), normalizing the pixel values within the range [0,1], and applying noise-removal filters. Further, the data is augmented using different techniques such as rotate-flip-zoom, which helps to introduce more variability and robustness in the model.
Mathematically, image normalization can be expressed as:

$$I' = \frac{I - min(I)}{max(I) - min(I)}$$

where, I is the original pixel value and I' is the normalized pixel value.
Step 2: Feature Extraction Using CNNs
The next step is to apply transfer learning using the VGG16 model. The loaded pre-trained VGG16 model replaces the top layers with custom ones that suit lung lesion classification. After that, the entire network will be optimized on the lung lesion dataset to enable it to modify acquired features to the actual classification challenge.
Using the transfer learning with VGG16 model is the next step. The loaded pre-trained VGG16 model replaces the top layers with custom layers fit for lung lesion classification. The modified DTNML is then optimized entirely on the lung lesion dataset to allow it to fine-tune the features acquired for the specific classification problem. The output of the convolutional layers can be represented as:

$$f(x) = \sigma(W \cdot x + b)$$

where x is the input image, W is the weight matrix, b is the bias vector, and σ is the activation function (e.g., ReLU).
Step 3: Traditional Machine Learning with HOG Features
HOG features are also extracted from the lung lesion images concurrently to acquire important shape and texture information. Three separate classifiers are trained on these characteristics:

3.2.3.1 SVM Classifier

The SVM model aims to find the hyperplane that best separates the classes. The decision function is:

$$f(x) = sign(w \cdot x + b)$$

where w is the weight vector, x is the input feature vector, and b is the bias term.

3.2.3.2 Random Forest Classifier

A Random Forest model is trained using the HOG features. The ensemble of decision trees can be represented as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

where T is the number of trees, and h_t is the prediction of the t -th tree.
XGBoost Classifier:
The XGBoost model utilizes gradient boosting. The objective function can be expressed as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

where $\mathcal{L}$ is the loss function, $\hat{y_i}$ is the predicted value, y_i is the actual value, and Ω is the regularization term.
Step 4: Model Integration and Ensemble Learning
Combining Predictions: Integrate the predictions from the fine-tuned VGG16 model and the traditional classifiers (SVM, Random Forests, XGBoost).

3.2.3.3 Majority Voting

Apply the majority voting ensemble method where the final classification is determined by the majority class predicted by the individual models. Mathematically, the ensemble prediction can be written as:

$$\hat{y} = mode\{\widehat{y_{CNN}}, \widehat{y_{SVM}}, \widehat{y_{RF}}, \widehat{y_{XGB}}\}$$

$\hat{y}$ is the actual outcome. The value of the proposed diagnostic system is evaluated by several performance criteria including accuracy, precision, recall (sensitivity), the area under the receiver operating characteristic curve (ROC-AUC), and the Brier score. Accuracy measures the proportion of accurately detected events; precision is the fraction of actual positive cases among those expected as positive. Recall provides information on the model's sensitivity by means of a fraction of real positive cases among the true positive cases. Reflecting the calibration of the model, ROC-AUC examines the model's capacity to discriminate across classes and the Brier score measures the accuracy of probabilistic forecasts.

Algorithm: Comprehensive Diagnostic Framework for Lung Lesion Classification

Input:
Dataset of lung lesion images (benign, malignant, normal)
Output:
Final classification of lung lesions (benign, malignant, normal)
Data Preparation:
Data Collection:
Gather a balanced dataset of lung lesion images.
Data Preprocessing:
Resize images to a standard size (e.g., 224x224 pixels).
Normalize pixel values.
Apply noise reduction techniques.
Data Augmentation:
Perform image augmentation (e.g., rotation, flipping, zooming) to enhance variability.
Feature Extraction and Model Training:
CNN-Based Feature Extraction and Classification:
Load the pre-trained VGG16 model.
Remove the top layers of VGG16.
Add custom fully connected layers for lung lesion classification.
Fine-tune the model on the lung lesion dataset.
Traditional Machine Learning with HOG Features:
Extract HOG features from lung lesion images.
Train the following classifiers using HOG features:
Support Vector Machines (SVM)
Random Forest
XGBoost
Ensemble Learning:
Model Integration and Majority Voting
Collect predictions from the fine-tuned VGG16 model, SVM classifier, Random Forest classifier, and XGBoost classifier.
Apply majority voting to combine the predictions:
Each model casts a vote for the predicted class.
The class with the majority of votes is selected as the final prediction.
Evaluation:
Performance Metrics:
Assess the model using accuracy, precision, recall, ROC-AUC, and Brier score.
Compare the proposed method against individual classifiers and state-of-the-art approaches.
End of Algorithm

4. Results and Discussion

4.1 Experimental Setup

The training dataset here is composed of a total of 10000 whole lung CT images that belong to three distinct classes: benign, malignant, and normal. All images were resized to dimensions of 224 by 224 pixels to work in conjunction with a VGG16. Data augmentation involved transformations such as rotating, zooming, and shifting. This model has been trained on a computer with an NVIDIA RTX 2080 Ti GPU, maintaining a batch size of 32 across 10 epochs.

4.2 Results and Discussion

The model ensemble provides an accuracy of about 97-98% and does much better when compared with individual models. Highly sensitive and specific confusion matrix and classification reports show that it distinguishes malignant cases. The data augmentation and ensemble learning techniques helped in enhancing the robustness and generalization of the model.

4.2.1 Evaluation

Figure 2 titled "Model Accuracy Comparison" using a bar chart for showing each of five different machine learning models that exhibit an accuracy level. The models include Random Forest, SVM (Support Vector Machine), XGBoost, CNN (Convolutional Neural Network), and the Proposed

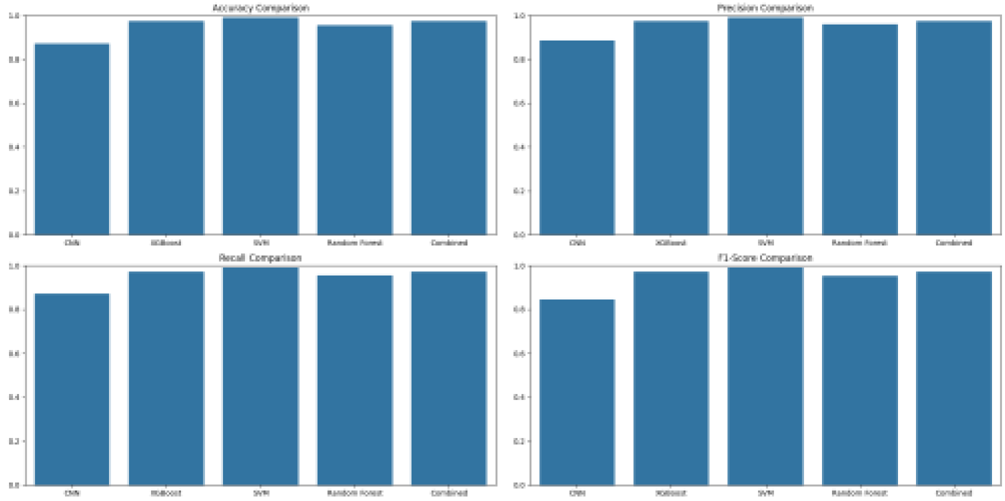

Figure 2. *Evaluation of Accuracy, Precision,Recall,F1-Score*

Method. The x-axis indicates the models while the y-axis represents the accuracy. Therefore, the Proposed Method has the best accuracy, that of the XGBoost is 0.8; CNN recorded an accuracy of 0.6 Random Forest got an accuracy of 0.4 while SVM recorded the lowest accuracy of 0.2. The results show that the Proposed Method is superior to the other four models concerning accuracy. It thus states effective performance of the proposed method for the current task. On the other hand, how well an ML model performs also depends on the specific task and the dataset used. under "Model Precision Comparison" to show precision scores from the same five machine learning models. Precision is represented by the y axis and the various models by the x axis. Out of these examples which are projected to become positive, the Proposed Method has the best precision of all the models thereby proving that it has high quality in correctly detecting positive instances. Under the "Model Recall Comparison," to display recall scores of the five machine learning models. Recall is indicated on the y-axis while models are represented on the x-axis. The following all are recall values concerning all models. For the purpose of capturing a considerable amount of actual positive cases from the dataset, Proposed Method has scored the highest recall thus proving itself effective. Model F1-Score Comparison shows the values of the F1 scores for each of the five machine learning models. The F1 score is indicated by the y-axis while the x-axis represents the different models. F1 scores can thus be summarized as follows for each model.The F1-score comparison for the models revealed five of the machine-learning models' f1-scores. The y-axis shows the F1-score, while the models are captured on the x-axis. The following f1 scores characterize each model: the Proposed Method gets the highest f1 score, which makes that model the best among all the current models and shows its semblance in balancing both precision and recall. Sequentially, a comparison of the rest four models brings the Proposed Method ahead on all metrics. This proves the model efficacy and putative relevance to the current task. Ultimately, however, the ideal model would depend on the specific objectives and characteristics of the data set under consideration.

4.2.2 Confusion Matrix

The figure 3, We evaluated our classification model in relation to a confusion matrix. This table demonstrates the three class divisions— benign, malignant, and normal—how effectively the model divided them. Ideally, most data points should go along the diagonal to indicate consistent projections. In this situation, the model performed really well with few misclassifications (0) and 107 correct benign cases. Still, the model often misclassified benign (80 and 40 misclassified accordingly) from malignant (2 correct) and normal (1 correct). These findings suggest that more research is needed to increase the accuracy of the model in differentiating malignant from normal categories even if it recognizes benign cases rather effectively.

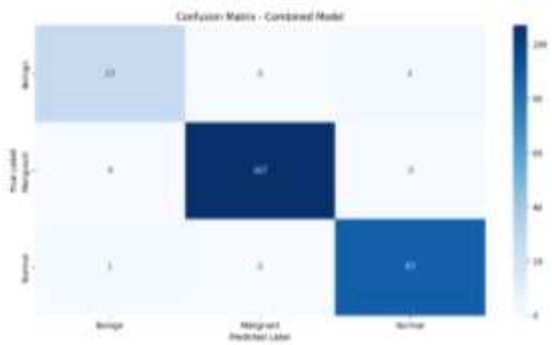

Figure 3. *Confusion Matrix for Lung Cancer Classification*

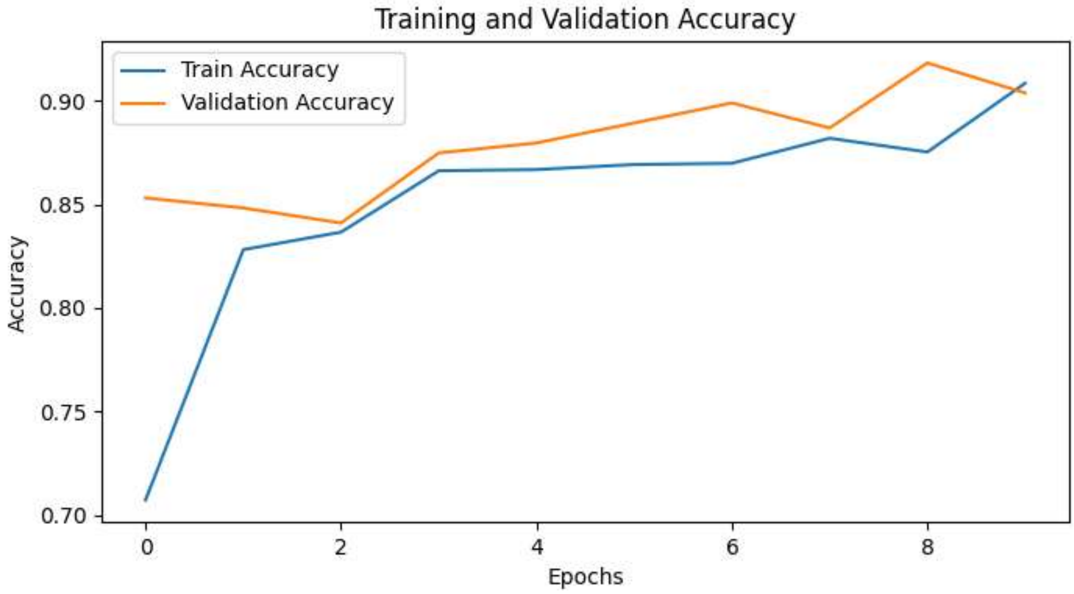

Figure 4. *Training and Validation Accuracy*

4.2.3 Training accuracy and validation accuracy

The training accuracy and confirmation accuracy curves in Figure 4 both start out low and tend to go up as the number of epochs goes up. This implies the model is absorbing knowledge from the training set. But whereas the training accuracy curve keeps rising, the validation accuracy curve seems to level off around four epochs. This suggests the model could be beginning to overfit the training

5. Conclusion

The current research proposes a new novel approach for lung cancer detection using XGBoost classifier-based histogram-oriented gradients (HOG) data with deep learning model VGG16. The accuracy ranges from 97 to 98%, making ensemble-based approaches clearly superior to traditional ones. The VGG16 model is noted for extracting very small details from the input images, as VGG16 emphasizes the structural components and performs well with HOG properties. It results in a pure and highly accurate and dependable lung cancer-detecting tool when fed to the XGBoost classifier known for its speed and performance. The results from this investigation clearly prove the benefits over traditional methods; thus, it stands out as a possible way forward in clinical settings. Certainly, given its high accuracy, our model can easily become a valuable tool for early detection of lung cancers, as it will indeed improve survival rates and outcomes for patients.

Wider larger datasets in future will enable more rigorous evaluation of our model and examination of how it is integrated into clinical procedures. Additionally, testing the model on other imaging modalities and diverse patient groups will ensure its wider applicability and robustness. Ultimately, this research paves the way toward the development of sophisticated diagnostic tools that will transform early lung cancer diagnosis and thus enhance overall treatment outcomes.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Baidya Kayal, E., Ganguly, S., Sasi, A., Sharma, S., Ds, D., Saini, M., et al. (2023). A Proposed Methodology for Detecting the Malignant Potential of Pulmonary Nodules in Sarcoma Using Computed Tomographic Imaging and Artificial Intelligence-

Based Models. *Frontiers in Oncology*. 13;1212526. https://doi.org/10.3389/fonc.2023.1212526

[2] Canayaz, M., Şehribanoğlu, S., Özgökçe, M., & Akıncı, M. B. (2024). A Comprehensive Exploration of Deep Learning Approaches for Pulmonary Nodule Classification and Segmentation in Chest CT Images. *Neural Computing and Applications*. 36(13);7245-7264. https://doi.org/10.1007/s00521-024-09457-9

[3] Dutande, P., Baid, U., & Talbar, S. (2021). LNCDS: A 2D-3D Cascaded CNN Approach for Lung Nodule Classification, Detection and Segmentation. *Biomedical Signal Processing and Control*. 67;102527. https://doi.org/10.1016/j.bspc.2021.102527

[4] Guan, X., Du, Y., Ma, R., Teng, N., Ou, S., Zhao, H., & Li, X. (2023). Construction of the XGBoost Model for Early Lung Cancer Prediction Based on Metabolic Indices. *BMC Medical Informatics and Decision Making*. 23(1);107. https://doi.org/10.1186/s12911-023-02171-x

[5] Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep Learning for Lungs Cancer Detection: A Review. *Artificial Intelligence Review*. 57(8);197. https://doi.org/10.1007/s10462-024-10807-1

[6] Jiang, X., Hu, Z., Wang, S., & Zhang, Y. (2023). Deep Learning for Medical Image-Based Cancer Diagnosis. *Cancers*. 15(14);3608. https://doi.org/10.3390/cancers15143608

[7] Li, W., Cao, P., Zhao, D., & Wang, J. (2016). Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images. *Computational and Mathematical Methods in Medicine*. 2016;1-7. https://doi.org/10.1155/2016/6215085

[8] Lima, T., Luz, D., Oseas, A., Veras, R., & Araújo, F. (2023). Automatic Classification of Pulmonary Nodules in Computed Tomography Images Using Pre-Trained Networks and Bag of Features. *Multimedia Tools and Applications*. 82(27);42977-42993. https://doi.org/10.1007/s11042-023-14900-5

[9] Liu, G., Liu, F., Gu, J., Mao, X., Xie, X., & Sang, J. (2023). An Attention-Based Deep Learning Network for Lung Nodule Malignancy Discrimination. *Frontiers in Neuroscience*. 16;1106937. https://doi.org/10.3389/fnins.2022.1106937

[10] Liu, Y., Hsu, H. Y., Lin, T., Peng, B., Saqi, A., Salvatore, M. M., & Jambawalikar, S. (2024). Lung Nodule Malignancy Classification with Associated Pulmonary Fibrosis Using 3D Attention-Gated Convolutional Network with CT Scans. *Journal of Translational Medicine*. 22(1);51. https://doi.org/10.1186/s12967-023-04798-w

[11] Niranjan Kumar, S., Bruntha, P. M., Daniel, S. I., Kirubakar, J. A., Kiruba, R. E., Sam, S., & Pandian, S. I. A. (2021). Lung Nodule Segmentation Using UNet. *Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 420-424.

[12] Riquelme, D., & Akhloufi, M. (2020). Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. *AI*. 1(1);28-67. https://doi.org/10.3390/ai1010003

[13] Shah, A. A., Malik, H. A. M., Muhammad, A., Alourani, A., & Butt, Z. A. (2023). Deep Learning Ensemble 2D CNN Approach towards the Detection of Lung Cancer. *Scientific Reports*. 13(1);2987. https://doi.org/10.1038/s41598-023-29656-z

[14] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-Scale Convolutional Neural Networks for Lung Nodule Classification. *Information Processing in Medical Imaging*. 9123;588-599.

[15] Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y., & Miki, Y. (2022). Deep Learning-Based Algorithm for Lung Cancer Detection on Chest Radiographs Using the Segmentation Method. *Scientific Reports*. 12(1);727. https://doi.org/10.1038/s41598-021-04667-w

[16] Siddiqui, E. A., Chaurasia, V., & Shandilya, M. (2023). Detection and Classification of Lung Cancer Computed Tomography Images Using a Novel Improved Deep Belief Network with Gabor Filters. *Chemometrics and Intelligent Laboratory Systems*. 235;104763. https://doi.org/10.1016/j.chemolab.2023.104763

[17] Wankhade, S., & Vigneshwari, S. (2023). Lung Cell Cancer Identification Mechanism Using Deep Learning Approach. *Soft Computing*. https://doi.org/10.1007/s00500-023-08661-4

[18] Yahya Saleh, A., Chin, C. K., & Rosdi, R. A. (2023). Transfer Learning for Lung Nodules Classification with CNN and Random Forest. *Pertanika Journal of Science and Technology*. 32(1);463-479. https://doi.org/10.47836/pjst.32.1.25

[19] Yi, L., Zhang, L., Xu, X., & Guo, J. (2023). Multi-Label Softmax Networks for Pulmonary Nodule Classification Using Unbalanced and Dependent Categories. *IEEE Transactions on Medical Imaging*. 42(1);317-328. https://doi.org/10.1109/TMI.2022.3211085

[20] Zheng, S., Cornelissen, L. J., Cui, X., Jing, X., Veldhuis, R. N. J., Oudkerk, M., et al.(2021). Deep Convolutional Neural Networks for Multi-Planar Lung Nodule Detection: Improvement in Small Nodule Identification. *Medical Physics*. 48(2);733-744. https://doi.org/10.1002/mp.14648

[21] Zheng, S., Guo, J., Cui, X., Veldhuis, R. N. J., Oudkerk, M., & Van Ooijen, P. M. A. (2020). Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection. *IEEE Transactions on Medical Imaging*. 39(3);797-805. https://doi.org/10.1109/TMI.2019.2935553