Journal of Soft Computing and Computer Applications

Manuscript 1011

Development of a Hybrid Methodology of Deep Learning and Machine Learning for Lung Nodule Detection in Medical Computed Tomography Images

Zaed S. Mahdi

Rana M. Zaki

Alaa Kadhim Farhan

Negar Majma

JOURNAL OF SOFT COMPUTING AND COMPUTER APPLICATIONS 2024;1:1011

ORIGINAL STUDY

Development of a Hybrid Methodology of Deep Learning and Machine Learning for Lung Nodule Detection in Medical Computed Tomography Images

Zaed S. Mahdi [a,*], Rana M. Zaki [b], Alaa Kadhim Farhan [b], Negar Majma [c]

[a] University of Technology – Iraq, Information Technology Center, Al-Sina'a St., Al-Wehda District, 10066 Baghdad, Iraq
[b] University of Technology – Iraq, Department of Computer Science, Al-Sina'a St., Al-Wehda District, 10066 Baghdad, Iraq
[c] Naghshejahan Higher Education Institute, Computer Science Department, Isfahan, Iran

ABSTRACT

Deep learning and machine learning play an important role in the medical field, helping doctors make accurate, fast and effective diagnosis. Despite the progress achieved in the use of modern technologies in detecting cancerous nodes, current studies still suffer from some challenges and limitations that must be addressed to obtain high efficiency in identifying cancerous nodes. These challenges include using image pre-processing, combining deep learning and machine learning techniques, and constantly adapting to clinical changes, in order to address this. A hybrid methodology has been proposed for detecting cancerous nodules in the lung in medical Computed Tomography (CT) images. It includes three stages, the first of which is pre-processing the images used, identifying nodes, balancing samples, and extracting features using Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) from the images, and the second is based on building a deep learning model consisting of Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) to extract the trained features and combine them with the features extracted by LBP and HOG to be input for the next stage. In the final stage, the improved eXtreme Gradient Boosting (XGBoost) machine learning classifier is built using the Particle Swarm Optimization (PSO) algorithm. The highest accuracy results of 99.3%, and ROC-AUC were obtained 99.8%. The proposed methodology has proven its efficiency in detecting cancerous nodes accurately using CT images.

Keywords: Hybrid learning methodology, Computed tomography, eXtreme Gradient Boosting, Optimization, Local Binary Pattern and Histogram Oriented Gradients

Received 20 October 2024; accepted 20 December 2024.
Available online 31 December 2024

* Corresponding author.
E-mail addresses: zaed.s.mahdi@uotechnology.edu.iq (Z. S. Mahdi), rana.m.zaki@uotechnology.edu.iq (R. M. Zaki), Alaa.k.farhan@uotechnology.edu.iq (A. K. Farhan), Majma@naghshejahan.ac.ir (N. Majma).

https://doi.org/10.70403/3008-1084.1011

1. Introduction

Lung cancer is the most common and fatal type of cancer in the world. It represents a major challenge for global health. Detecting cancerous lung nodules is one of the most important steps in early diagnosis of lung cancer. Early diagnosis by detecting cancerous nodes in the lung increases the chances of treatment. Nodules can be detected using medical imaging techniques such as X-rays and Computed Tomography (CT), which require a long time and high expertise from doctors to analyze the medical images [1]. With the development in the field of technology, deep learning, machine learning, and artificial intelligence techniques are important tools and have proven their worth in the field of medicine and analysis of medical data [2] and images, especially in the field of lung diseases and cancer diagnosis [3–5]. Deep learning helps understand medical images and train the model to detect cancerous nodes, and machine learning can be used to classify the types of these nodes. Artificial intelligence helps identify and analyze patterns in this data [6]. Recently, many systems have been designed for early detection of cancerous diseases in the lung, using different techniques, some of which rely on deep learning and others rely on machine learning, due to their effectiveness in analyzing and evaluating diseases [7].

Current studies to detect cancerous diseases in the lung face many challenges and limitations, which include many technical aspects [8]. The medical data available for training is limited in terms of quantity and quality, and it is also necessary to deal with data that is balanced in terms of training samples [9]. Other challenges include pre-processing the medical images used, methods for extracting useful features from them, normalizing the data, the interpretability of the model to be well clear to doctors, and adapting to changes, continuing clinical practices and continuous updating, and finally choosing the appropriate techniques for this field of deep learning and machine learning [10–12].

The main goal of this research paper, a hybrid methodology was proposed to improve the accuracy of detecting lung cancer nodes in CT medical images, based on deep learning, machine learning, and artificial intelligence techniques. First, the data is analyzed and processed, extracting node coordinates, converting them to voxel coordinates, identifying classes, learning data series and extracting features using Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) from medical images. Then, a neural network model consisting of Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) was built, trained, and the trained features were extracted, using early stopping and learning rate scheduling to avoid over-adaptation, and then merging the trained features with the features extracted from the images (LBP and HOG). The combined features are then fed into the XGBoost classifier augmented by Particle Swarm Optimization (PSO). To evaluate the model's performance, the LUng Nodule Analysis (LUNA16) data set was used, which contains medical CT, and the accuracy, Area Under the Curve (AUC(, and confusion matrix were computed for each stage of the work.

In the proposed methodology, all the challenges and limitations faced by current studies were addressed, in terms of the accuracy of the data used, balancing samples for training, using more than one technique, taking advantage of the advantages of each technique, clarity and ease in showing the results to doctors, and also the possibility of development in the future to keep pace with the development of diseases.

The contributions of the research paper are as follows:

- Developing a hybrid methodology of deep learning techniques, advanced machine learning, artificial intelligence techniques, and image processing techniques, which have proven effective in improving the overall performance of the proposed methodology.

- Providing a reproducible and practical framework for clinical testing.
- Visual display of diagnosis, where functions have been developed to display the prediction results of existing nodes, which helps doctors in rapid diagnosis.

The remainder of this paper is presented as follows:

The second section presents work related to cancer node detection models. The requirements and an explanation of the techniques used in the proposed methodology (image processing, deep learning and machine learning) are presented in the third section. The fourth section presents a detailed explanation of how the proposed hybrid methodology works. Details of the performance evaluation and dataset are presented in fifth Section. As for the sixth section, a detailed explanation of the results of the performance evaluation was made, comparisons were made between the results of each stage, the results of the methodology were also compared with related work, and the results of the performance evaluation were discussed. Finally, in Section Seven, the conclusion and suggested future works are presented.

2. Related works

In this section, previous literature on cancer node detection systems, their characteristics and limitations will be analyzed.

The authors in [13] proposed a study to detect cancerous nodules in the lung, using computed tomography. It relies on deep learning algorithms. They also used Boosting algorithms to improve the accuracy of the final predictions, and to increase the diversity of data, they used data augmentation algorithm. The model's performance was evaluated on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) CT image dataset, and the model's accuracy was 84.8%. In [14], the authors presented a hybrid model for lung nodule detection, based on deep learning and machine learning algorithms. They used the CNN algorithm to identify patterns in CT images. To extract features, they used Capsule Network. In classifying diseases, they used the Support Vector Machines (SVM) algorithm, and adjusted the parameters for both of CNN and SVM. In evaluating the model's performance, they used the LIDC-IDRI data set for CT images, and the accuracy of the hybrid model was 94%. The authors in [15], proposed a model for detecting cancer in the lung based on CT images. The model was based on CNN algorithms with performance optimization algorithms, and the best model was combining the AlexNet model with the Stochastic Gradient Descent)SGD(algorithm, and in evaluating the performance they used the LUNA16 data set. The accuracy was 97.42%. The authors in [16] proposed a deep learning model based on integrating more than one CNN to detect cancerous nodes in the lung, by deep ensemble 2D CNN has different layers, kernels and pooling techniques. In performance evaluation and training, they used the LUNA16 dataset, which contains medical CT, and the accuracy of the combined model was 95%. The authors in [17] proposed a Deep Learning Convolutional Neural Network (DLCNN) methodology to detect cancerous nodes in the lung based on the CNN algorithm and machine learning to predict nodes, as well as processing medical CT images. The simulation was conducted on the LUNA16 data set, and the accuracy of the model was 92.81%. The authors presented in [18] a hybrid model of neural networks and machine learning to detect cancer nodes, where the neural networks consist of 7 convolutional networks, three pooling layers, and two connected layers, and through them, features are extracted to be used later in the SVM classifier, and they used the LUNA 16 data set. To evaluate the performance, the accuracy of the model was 97.64%. Authors in [19] presented a deep learning model for detecting lung cancerous nodes in medical CT, which is based on the 3D-VNet architecture for fine segmentation and the 3D Residual Network)3D-ResNet (architecture. In evaluating the performance, the LUNA16 dataset was used, and the accuracy of the model was 99.2%. The authors in [20] proposed a model based on a CNN under the name "You only look once" to detect cancerous nodes in the lung. They also improved the weights of the CNN through biogeography-based optimization with extinction and evolution, and they also pre-processed the data. Two data sets were used to evaluate the model's performance, namely LIDC-IDRI and LUNA 16. The accuracy results were 90.01% and 92%. The authors in [21], used medical CT images to detect cancerous nodes in the lung by proposing a methodology based on deep learning, and the model was a hybrid between RU-Net and R2U-Net, and LSTM networks with the use of the Visual Geometry Group – 16)VGG16(architecture., and merged them together to obtain a final model for identifying nodes. To evaluate the model's performance, they used the LUNA16 data set, and the model's accuracy was 90%. The authors in [22], proposed an Adaptive Large Kernel)ALK U-Net (model, based on several algorithms. In the beginning, they used the ALK U-Net algorithm to extract features, and then they used spatial and channel attention, and integrated them with the network to enhance the distribution of weights, after which they divided the model. It is divided into two parts using the U-Net architecture algorithm for encryption and decryption. To evaluate the model's performance, they used LUNA-16 data set, and the similarity coefficient Dice Similarity Coefficient (DSC) was 91.57%.

3. Materials

The use of modern technologies of deep learning and machine learning in the medical field puts researchers in front of great challenges, especially after began to rely on them in diagnosing and identifying incurable diseases. It has become necessary for the diagnosis to be with high accuracy since it concerns human life, and among its uses is the detection of cancerous diseases in lung and node identification using medical CT images [23].

3.1. LUNA16 dataset

LUNA16 is a medical dataset containing CT images of the lung, including signs of pulmonary nodules [24]. Taken from several patients, it is used for research and diagnosis of cancerous lung diseases. It contains other information about the images, such as the image ID, voxel size, and the number of slices in each image [25]. It contains patient data, and also contains the node coordinates of the images (x, y, z) and their radius. This data set consists of 1040 CT images [26], Fig. 1 and Fig. 2 shows samples of the dataset.

In Fig. 2, the number of samples in the dataset used is shown. The number of positive samples is 818 and the number of negative samples is 222.

3.2. Image processing techniques

Image processing techniques play an essential role in working with images in deep and machine learning models [27], as it is necessary to create filters for images, remove noise, and identify patterns within images, in addition to the process of normalizing images to be easy to work with within models [28].

3.2.1. Median filtering

It is one of the image processing techniques used to remove noise from images [29]. It is considered effective in removing pulsed noise or salt and pepper noise without affecting

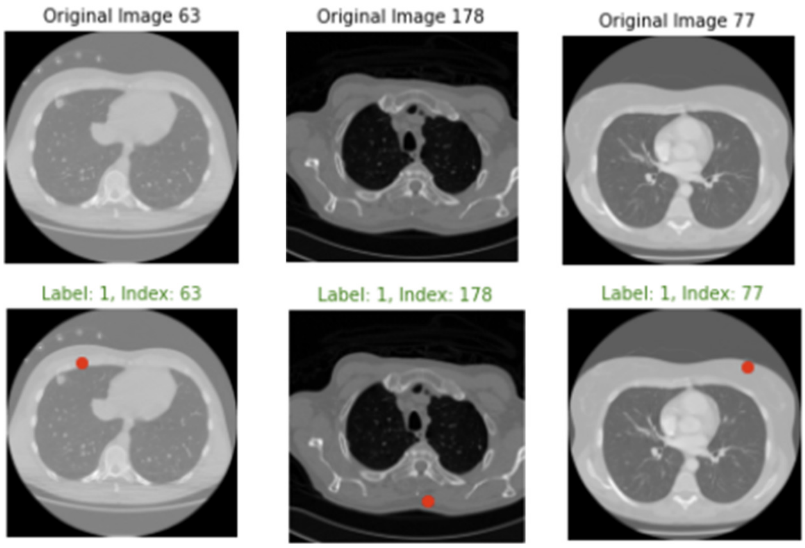

Fig. 1. Samples in the LUNA 16 dataset and dropping nodes onto the images.

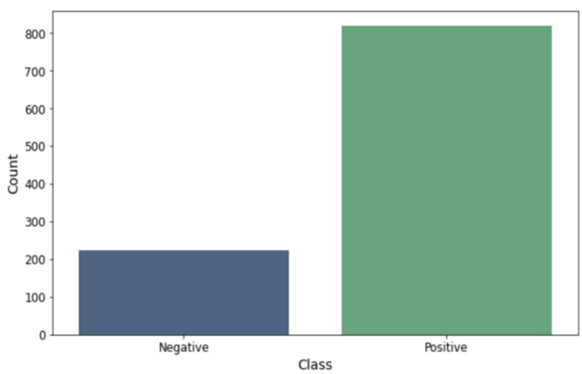

Fig. 2. Distribution of Samples in dataset LUNA 16.

the edges. It is widely used in medical images [30]. The way it works depends on choosing a small window with specific dimensions to move it across each pixel. The values of the pixels are then collected within the window surrounding the target pixel, after which the median value is calculated, after which the window is moved and the steps are repeated for each pixel in the image [31].

3.2.2. Resizing and normalization

Resizing is the process of adjusting the dimensions of images to fixed dimensions, regardless of the size of the original images. This is done through two-dimensional or three-dimensional interpolation algorithms, and the benefit of them is to facilitate their processing by neural networks, which helps in accelerating training and ensuring that all images are processed in the same way [32]. Normalization in which the pixel values of images are changed to (0 and 1), each pixel value is divided into the maximum possible value of the pixel, and this technique contributes to facilitating training by reducing the discrepancy between the values, which leads to improving the work of the neural network and helps in better stability of the weights during the process, training and giving the best results [33].

3.3. Feature extraction tools

It is an important part of the image processing process, as extracting features from images plays a major role in enhancing other features, which contributes to improving the performance of deep learning models and providing accurate and effective results [34, 35].

3.3.1. Local binary pattern

It is one of the methods used to extract relevant features from images, which is the most common in analyzing patterns within images. It is characterized by simplicity and speed of calculation. Its working method is based on comparing each pixel in the image with the other pixels surrounding it, and a unique binary pattern is created and calculated, decimal value and building a histogram [36].

3.3.2. Histogram of oriented gradients

It is the extraction of features from the edges of images, as it can recognize shapes and objects and is based on analyzing the trends of gradients in the image, and contributes to improving the performance of training deep learning models [37]. Its working principle is based on calculating the image gradients, dividing the image into small cells, creating a gradient histogram, normalizing the histograms and combining the histograms from all the blocks to form the final HOG feature of the image [38].

3.4. Deep learning models

They are neural networks that are used to analyze images, identify patterns, and obtain well-trained features that help in classifying cancerous nodes [39].

3.4.1. Convolutional neural network

It is considered one of the most commonly used tools for recognizing image patterns and extracting relevant features [40]. It includes several convolutional and pooling layers, where small filters are passed through the image and then weights are determined. Each layer can contain many filters [41]. The method of determining features is done through convolutional layers, activation layers, pooling layers, projection layers and dense layers. Through this structure, the neural network can determine the features of the trained images [42, 43].

3.4.2. Long-short-term memory

It is a type of recurrent neural network, designed to process sequences of information with high efficiency, and is used in image analysis and feature extraction by learning from context-dependent temporal patterns [44]. It includes input gate, forget gate, and output gate. The output from LSTM is passed to other layers to achieve the final task, such as classification or prediction [45, 46].

3.5. eXtreme gradient boosting

eXtreme Gradient Boosting (XGBoost) is used to build boosted gradient models effectively, and is distinguished by its ability to provide high-performance and speedy models. The basis of its currency is based on boosted gradient and tree of models algorithms that combine many weak trees to form a strong model. The best parameters are determined (learning rate, max depth, n estimators and subsample) to obtain an effective model [47, 48].

3.6. Particle swarm optimization algorithm

It is a stochastic optimization technique whose work is inspired by the behavior of a flock of birds. PSO is used to improve cost functions by updating the location and speed of particles in multi-dimensional space [49]. It initializes a number of particles in random space, and each particle represents a possible solution, and updates the positions and velocities based on the best location by the particle itself, and then the objective function is calculated. It can be used with models to find the best parameters and thus obtain the best model performance [50].

4. Proposed methodology

In this section, a detailed presentation of the proposed hybrid methodology will be presented to detect cancerous nodes in the lung and as shown in Fig. 3. In the first stage, data is loaded from a Comma-Separated Values)CSV(file that contains node coordinates and patient information, as well as loading CT images, processing medical images, projecting nodes coordinates onto the images, and performing a balancing act for samples that will be used to ensure good training of the model, LBP and HOG features are then extracted from the images for later use. In the second stage, a hybrid deep learning model is built, which consists of CNN and LSTM, and trained on samples, to obtain trained features, so that the trained features are combined with the LBP and HOG features extracted from the images to form an integrated feature set, i.e., feature fusion. In the third stage, the XGBoost classifier is built and the parameters are determined using the PSO algorithm, to obtain an improved classifier. After dividing the data into training data and test data, the improved classifier is trained to classify the nodes found in the CT images as positive or negative. The model's performance is then evaluated and metrics are calculated.

4.1. Data pre-processing

In data processing, the data containing the node coordinates and nominations data is first read, and the relevant features are extracted, namely the coordinates, categories, and string identifier, and the coordinates are converted into voxel coordinates using the origins and spacing extracted from the metadata, and then the metadata is read for CT images, and to improve the images, the median filter is used to remove noise from the images, and then the images are resized to fixed dimensions (128×128 pixels), To facilitate processing in the deep learning model. After that, the slices containing the cancerous nodes are extracted, and the coordinates of the nodes are projected onto the images, as shown in the Fig. 4. After that, a balance is made for the positive and negative samples to ensure good training, and to ensure the stability of training in the model, Also normalize the images to values (between 0 and 1) which also helps in back propagation in neural networks, At the end of data processing, LBP and HOG features are extracted from the images to be used later in training. This helps to enhance the trained features and obtain more information, which will help increase the accuracy of classification.

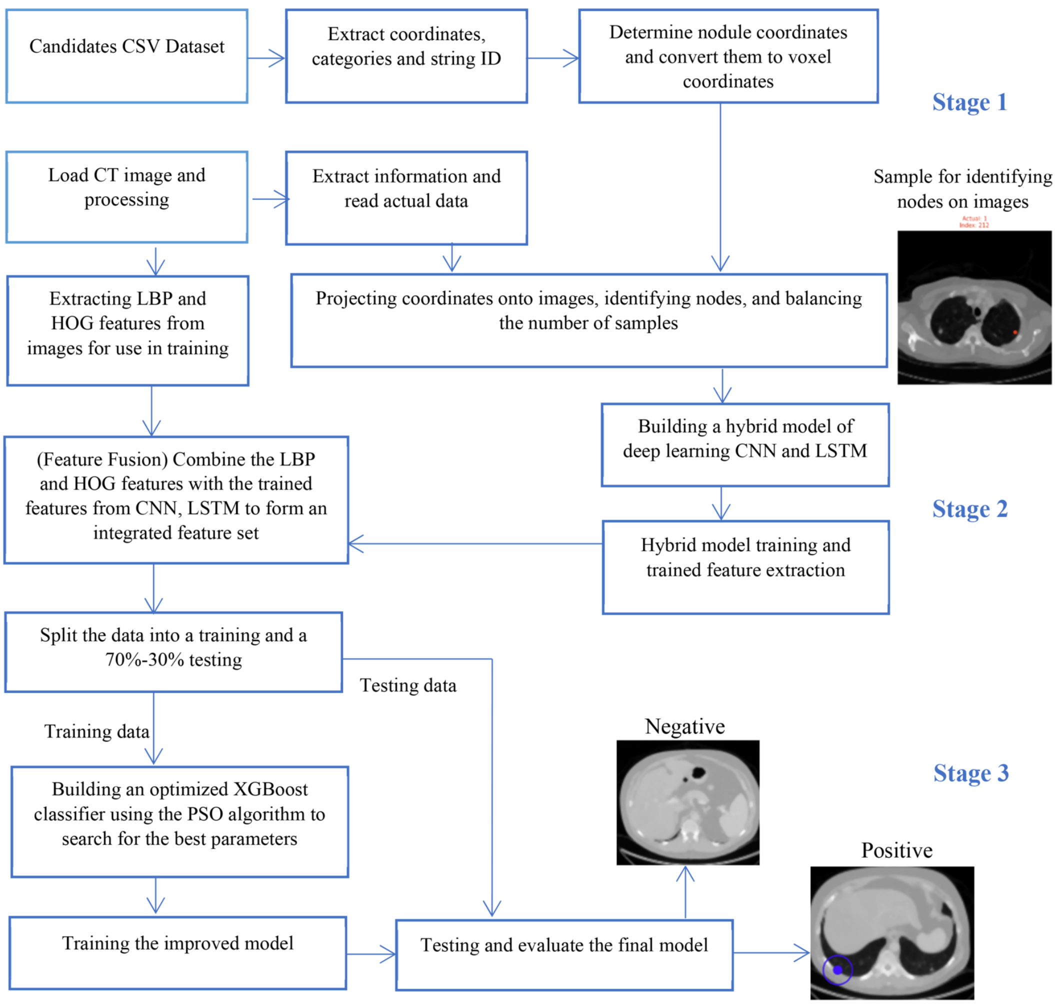

Fig. 3. Structure of the proposed methodology.

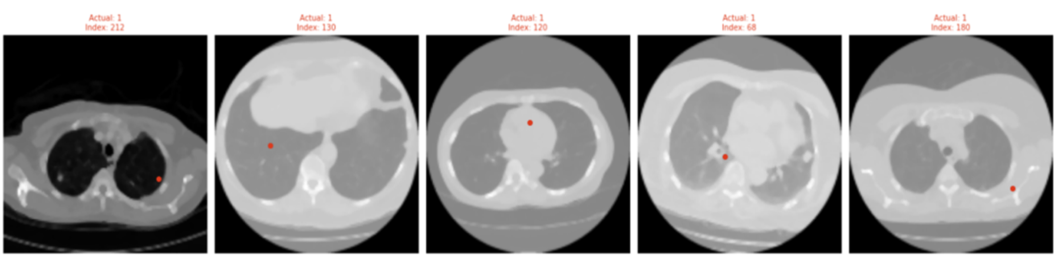

Fig. 4. Samples of projecting nodes onto images.

4.2. Building a hybrid deep learning model

The hybrid deep learning model includes the use of CNN layers and LSTM layers to process medical images and extract the trained features to be used later. CNN layers include 3 Conv2D, 3 MaxPooling2D, 3 Dropout layers, 3 Batch Normalization layers, and 1 Flatten

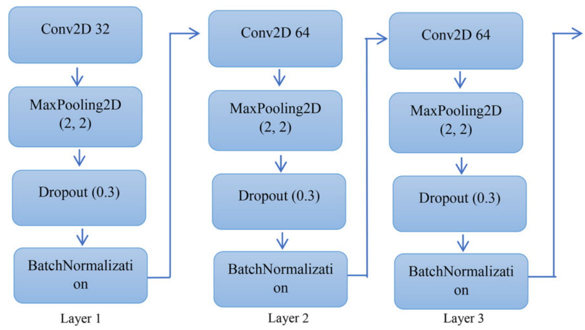

Fig. 5. Structure of the CNN algorithm used.

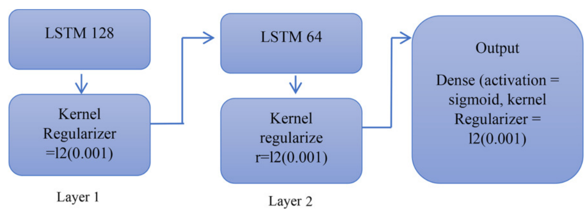

Fig. 6. Structure of the LSTM algorithm used.

layer, as shown in Fig. 5. Two layers for LSTM and one output layer (Dense Layer), as shown in Fig. 6. In Conv2D layers, Filters are used (Kernel size, activation function: ReLU, L2 regularization), including LSTM layer (units, return sequences, L2 regularization).

The two networks are combined to analyze the images. The CNN extracts spatial features from each image using the time distributed layer, allowing each image to be analyzed independently. The image features are then sent to the LSTM network to analyze temporal and sequential patterns in the image data. Finally, the data is classified using the Dense layer with Sigmoid activation function to produce a binary result (positive or negative). This combination improves the accuracy of the model by leveraging the power of both networks.

After completing the construction of the hybrid model for deep learning, the model is trained and early stopping and learning rate scheduler are used to ensure stability, improve training, and prevent over fitting and get the features trained used and ready. To enhance the model's performance and improve its ability to classify images, the feature fusion stage comes, which is to merge the features trained from the CNN and LSTM model with the LBP and HOG features, and form a comprehensive set of features to be used as input to the XGBoost classifier.

As for the hyper-parameters, only the following were used: learning rate = 0.0001, loss function was the binary cross entropy, Adam optimizer, batch size = 64, epochs = 100 with the use of early stopping, class weight to balance the effect of negative and positive samples and learning rate scheduler was used to reduce the learning rate ensure that overfitting does not occur during training

4.3. Construction of eXtreme gradient boosting classifier optimized by particle swarm optimization

In this stage, the features that were combined in the previous stage, extracted by the CNN and LSTM algorithms, and the LBP and HOG features will be used. All of these features are used in one matrix as inputs to the XGBoost model.

In the beginning, the data is divided into training data and test data in a ratio of 70:30, and then comes the final stage of the hybrid model, and the XGBoost classifier is built and the PSO algorithm is used to improve the model and determine the best set of parameters for the model by initializing particles, updating particles, calculating objective function and velocity update and learning rate from 0.001 to 0.1, max depth from 3 to 10 and subsample from 0.5 to 1.0, and then the model is trained on the training data set.

4.4. Testing and evaluation of the hybrid model

The goal of this stage is to evaluate the performance of the model in detecting cancerous nodules in the lung. At this stage, a previously invisible test set for the model will be used, which will help understand the model's work on new data, and the extent of its accuracy and effectiveness on new data. The confusion matrix, accuracy, precision, recall, and AUC curve will be computed and displayed.

5. Performance evaluation details

In this section, details of the performance evaluation of the proposed hybrid methodology for detecting cancerous nodes in the lung will be presented. The LUNA16 dataset was used, which contains medical CT images of the chest, and includes a set of positive and negative samples for cancerous nodes in the lung. The proposed methodology was applied to a workspace that has the following specifications: a 64 bit operating system, Intel Core i5-12860G CPU @2.8 GHz, 16 GB RAM, running on Windows 11, Python 3.10.4, and utilizing Jupyter Notebook.

In training and testing the proposed methodology for detecting cancerous nodes, the LUNA 16 data set was used. The total number of samples was 1,040, the number of positive samples was 818, and the number of negative samples was 222, as shown in Fig. 7. Then, LBP and HOG features are extracted from the images to enhance the features of the deep learning model. In training the deep learning model based on CNN and LSTM, all samples were used to obtain trained features, and combined with features extracted by LBP and HOG to form a final feature set used in the final classification model, as shown in Figure

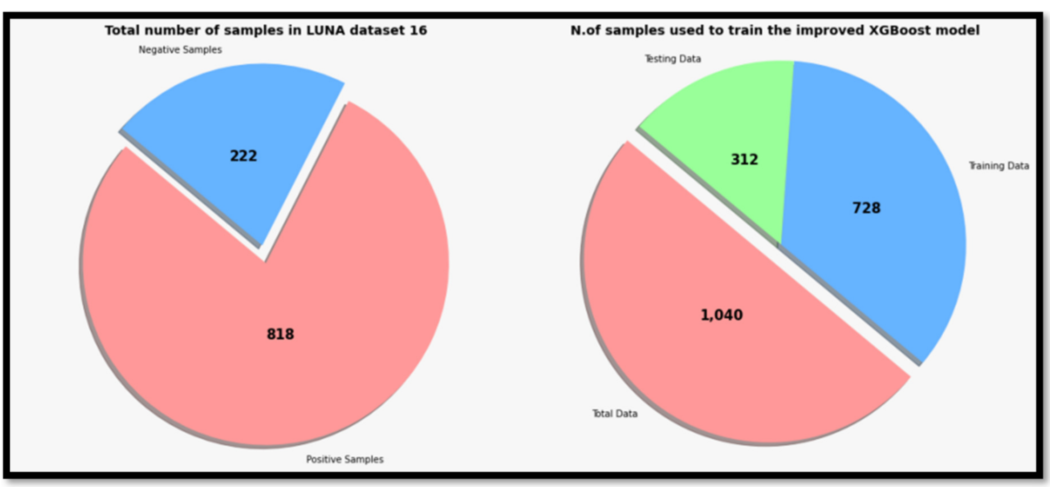

Fig. 7. Number of samples in the data set for each stage.

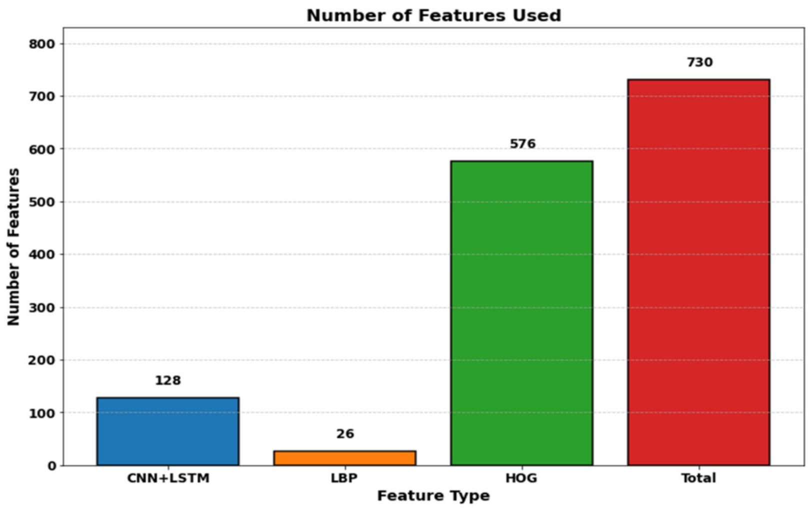

Fig. 8. Number of features.

7. In training the improved XGBoost model using PSO, the number of training data set is 728, and the test data is 31, as shown in Fig. 7, and In Fig. 8 the number of features for each stage is shown Fig. 7.

6. Performance evaluation results

This section presents the results of the performance evaluation of the proposed hybrid methodology for detecting cancerous nodes in the lung, and through the data set used for training, performance evaluation metrics will be calculated for each stage of the work, which are the confusion matrix, accuracy, recall and the AUC.

To evaluate the performance of the proposed hybrid methodology for detecting cancerous nodes in the lung, the LUNA16 dataset was used. The data was read and the medical CT images were processed, the samples used in the model were balanced, and features were extracted from the images by using LBP and HOG. Then a deep learning model, i.e., CNN and LSTM is trained to extract the trained features and combine them with the features extracted from the images to form an integrated feature set.

Fig. 9. Shows the confusion matrix of the deep learning model, i.e., CNN and LSTM. It shows the accuracy of the model in training and the decrease in cases of misdiagnosis. The accuracy of the model was 97.02%.

In Fig. 10, the training accuracy of the deep learning model is shown after stopping training in 52 epochs, where it continued to increase until it reached 97.02%, and it also shows the validation accuracy with test data, which reached 96.90%. The two values indicate that the model learns patterns effectively and also its ability to generalize well to new data.

After obtaining the trained features from the deep learning model and combining them with the features extracted by LBP and HOG, they are used as input to the XGBoost classifier improved by the PSO algorithm. The final model is trained and its ability to classify nodes found in medical CT images is tested.

Fig. 11. Shows the confusion matrix of the final hybrid model. The performance of the model is effective and has the ability to identify nodes well, it can be observe only two cases of false positive predictions and no false negative predictions. Fig. 12 shows how

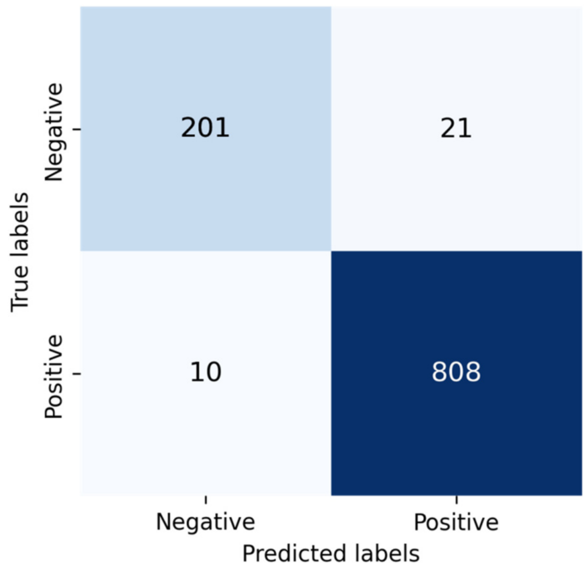

Fig. 9. Confusion matrix for deep learning model.

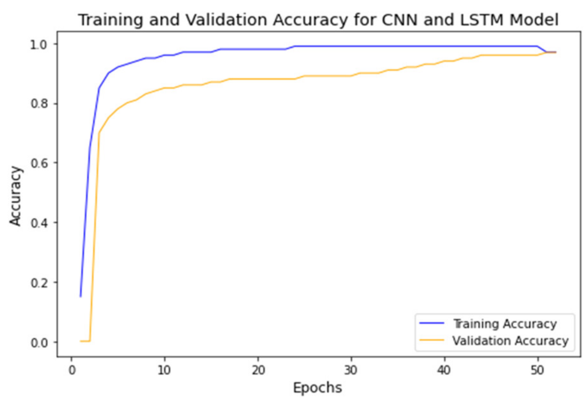

Fig. 10. The training and validation accuracy values for the deep learning model.

the images are displayed after predicting the nodes, where the node is highlighted in blue. The accuracy of the final hybrid model is 99.3%, the precision is 99.1%, recall 1.0%, error rate 0.064% and the cross-validation accuracy with test data was 99.1%. Fig. 12 shows the node prediction results, engraving them in blue, and enlarging the location of the node so that it is clear to see.

In order to ensure the efficiency of the proposed methodology, a comparison was made to detect nodes using deep learning algorithms only with the proposed hybrid methodology. Table 1 displays the accuracy results for the two stages of the proposed methodology. An

Table 1. Comparison between the results of the first and second stages.

Stage	Model	Accuracy	Recall	Precision	AUC	Error rate
Stage one	CNN and LSTM	97.03%	98.77%	97.46%	99.4%	0.298%
Stage two	XGBoost	99.3%	1.0%	99.1%	99.8%	0.064%

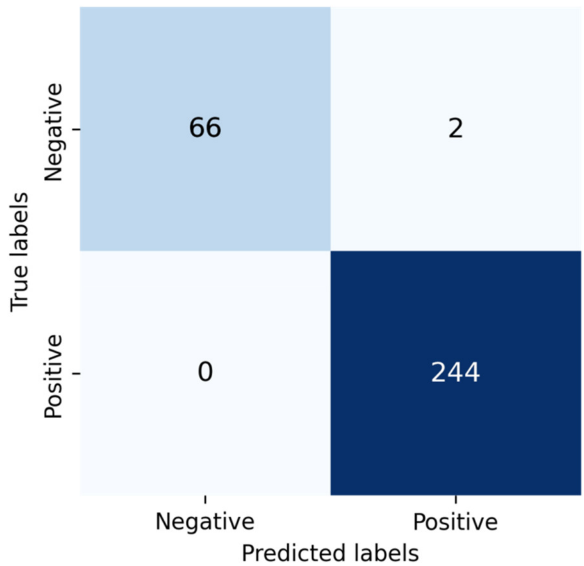

Fig. 11. Confusion matrix for final model.

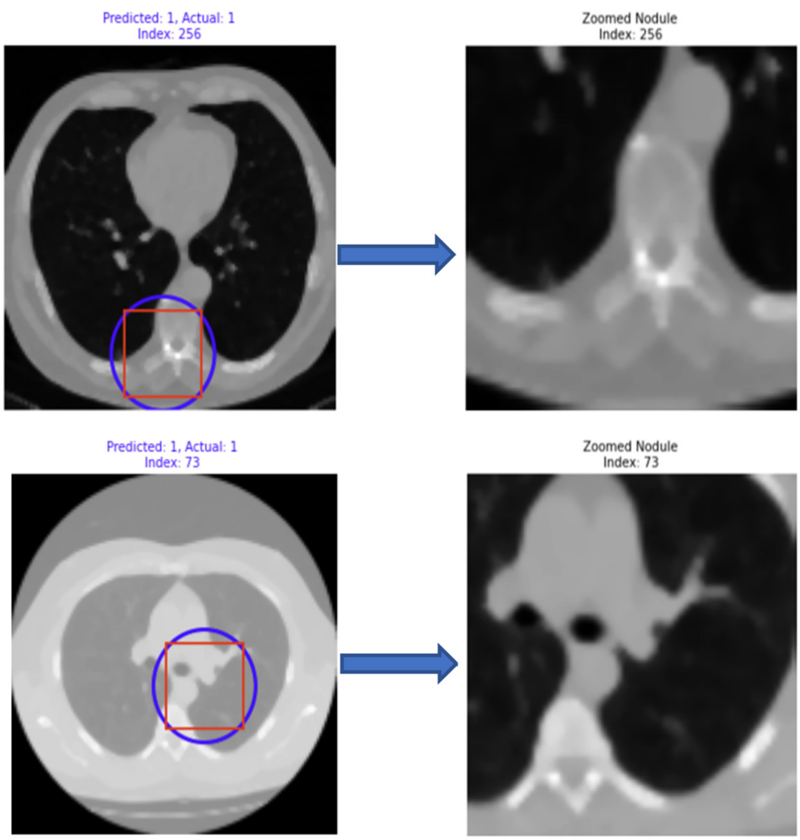

Fig. 12. A sample of prediction results zooming in on the node location.

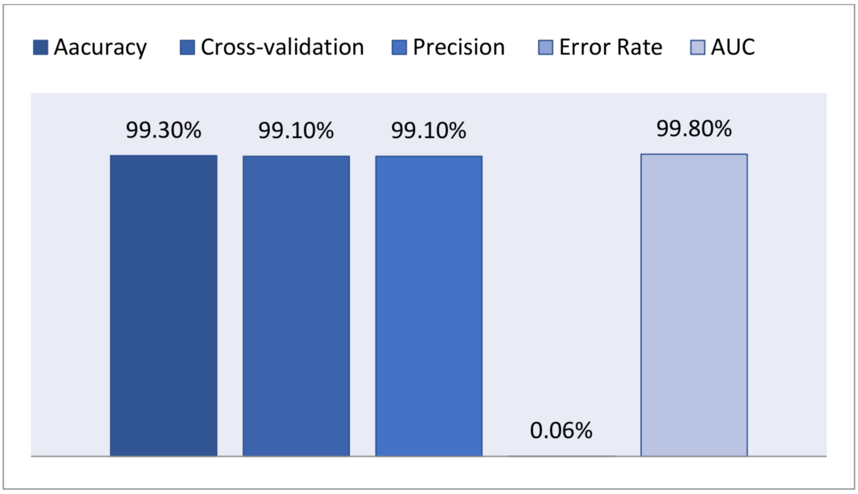

Fig. 13. Final results of the proposed model.

Table 2. Comparing the proposed hybrid methodology with related work.

Ref	Techniques used	Accuracy	Recall	Precision
[15]	AlexNet model with the SGD algorithm	97.42%	–	–
[16]	Deep Ensemble 2D CNN	95.6%	80.1%	93.7%
[17]	DLCNN	92.8%	92.3%	91.8%
[18]	CNN and SVM	97.6%	96.3%	99.1%
[19]	3D-VNet and 3D-ResNet	99.2%	–	–
[20]	CNN and optimization with extinction and evolution	92%	–	91.8%
[21]	RNN, LSTM and VGG16	90%	–	–
	Proposed hybrid methodology	99.3%	1.0%	99.1

increase in model accuracy appears after applying the second stage of the methodology. In Fig. 13, the final results of the proposed model are shown, proving the superiority of the proposed methodology in detecting cancerous nodules in the lung.

6.1. Comparison of the proposed hybrid methodology with related works

In order to evaluate the performance results of the proposed methodology, a comparison between the proposed hybrid methodology and number of current studies to detect cancerous nodes in the lung has been made, which use the same data set that was used in the proposed hybrid methodology, which is LUNA16.

Table 2 shows the comparison of the proposed hybrid methodology for the LUNA16 dataset. The results presented show the superiority of the proposed hybrid methodology. The accuracy reached 99.3%, and this reflects its effectiveness in detecting cancerous nodes in the lung through medical CT images.

6.2. Discussion

After training and testing the proposed methodology on the LUNA16 data set, to detect cancerous nodes in the lung using medical CT images, the effectiveness of the proposed hybrid methodology was proven through the results obtained, where in the first stage, which is a deep learning model, i.e., CNN and LSTM, it was accuracy of 97.03% was

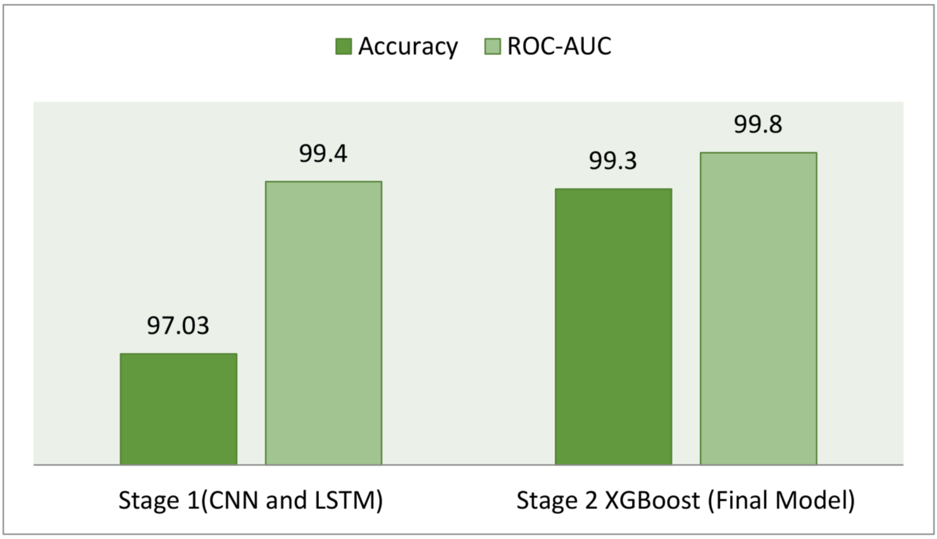

Fig. 14. Results of the proposed methodology by the datasets used.

obtained for the trained features. In the second stage, the trained features from the first stage were used as input to the improved XGBoost classifier using the PSO algorithm, which finds the best parameters. An accuracy of 99.3% was obtained for the final model. The results of the confusion matrix were very good in identifying nodes. Fig. 14 shows the results of the work stages.

In this study, many challenges were faced, including how to choose the appropriate data set, process images, and how to correctly identify nodes in images. Image processing steps, such as removing noise, filters, normalization, balancing, and extracting features from images, were absolutely necessary to obtain good results, as the work required the utmost precision to ensure the accuracy of the methodology. And also the method of integrating deep learning models, i.e., CNN and LSTM and avoiding complexity and training and obtaining trained features, where faced computational challenges, to avoid exposure to excessive training and maintain high accuracy through fine-tuning of parameters and applying techniques such as early stopping. Other challenges include choosing a suitable classifier for the work and how to choose its parameters using the PSO algorithm. As for the limitations of the proposed methodology, despite the good results obtained on the selected data set, the process of generalizing it to images and other data sets requires more work and research, as well as the process of testing the methodology clinically, and it is also necessary for continuous updates to be done. The methodology is due to the development taking place in the field of lung diseases. It is worth noting that this study is of great importance in designing a model to detect cancerous nodules in the lung through medical CT images, and also in assisting doctors in the diagnosis process.

The limitations facing the dataset used are the number of positive and negative samples, which can be overcome by balancing the data. Another limitation is the noise and lack of clarity present in some images, which results from the variety of devices in taking images, and which is treated through image filtering and processing techniques. These limitations lead to biases in training and generalization and also affect the results. Therefore, more than one technique was used to ensure that these limitations are overcome and good training is obtained for the proposed model.

Many challenges were faced in this study, including concern about over-adaptation, so several techniques were used, including early stopping, learning rate scheduling, which helped to overcome over-adaptation. Another challenge is the possibility of interpreting the results of the model to doctors, so the study of the model was facilitated and the results were presented so that they were well understood. Regarding implementation in the clinical environment, this will be part of future work to develop the proposed model.

7. Conclusion and future work

In the current study, a hybrid methodology between deep learning and machine learning was proposed to detect cancerous nodes in the lung. Deep learning consists of CNN and LSTM to extract the trained features and combine them with the features extracted from the images by using LBP and HOG, so that the set of features have been complete and use them. As input to the final classifier, the XGBoost classifier was used in machine learning, improved by using the PSO algorithm to find the appropriate parameters, and trained using the features that were previously combined. To evaluate the performance of the method, the LUNA16 dataset was used for medical CT images. The highest accuracy obtained was 99.3%, and the cross-validation accuracy was 99.1%. The proposed methodology addressed the challenges and limitations encountered in current studies, including the science of processing medical images, selecting appropriate data, and choosing hybrid methodologies to improve disease diagnosis models. In the end, this methodology proved to significantly improve the accuracy of the final model. Despite the challenges related to model complexity and time consumption. As for future work, it is important to advise the fellow researchers to use the proposed methodology with a new data set and conduct a clinical trial to adopt the methodology to detect lung cancer nodules.

Acknowledgement

The researchers extend their sincere thanks and appreciation to everyone who contributed to the completion of this research.

Conflicts of interest

The authors declare that there are no conflicts of interest with respect to the publication of this paper.

Authors' contributions

The authors confirm contribution to the paper as follows: study conception, design, data collection, analysis and interpretation of results: Zaed S. Mahdi, Rana M. Zaki; data collection, analysis and interpretation of results. Draft manuscript preparation, reviewed the results and approved the final version of the manuscript: Alaa Kadhim Farhan and Negar Majma.

Data availability

The data in this study are publicly available and can be accessed at https://www.kaggle.com/datasets/avc0706/luna16.

References

1. H. T. Gayap and M. A. Akhloufi, “Deep machine learning for medical diagnosis, application to lung cancer detection: a review,” *BioMedInformatics*, vol. 4, no. 1, pp. 236–284, Jan. 2024, doi: 10.3390/biomedinformatics4010015.
2. J. Latif, C. Xiao, A. Imran, and S. Tu, “Medical imaging using machine learning and deep learning algorithms: a review,” in *Proc. IEEE 2019 2nd Int. Conf. on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, pp. 1–5, doi: 10.1109/ICOMET.2019.8673502.
3. S. H. Hosseini, R. Monsefi, and S. Shadroo, “Deep learning applications for lung cancer diagnosis: a systematic review,” *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14305–14335, July 2023, doi: 10.1007/s11042-023-16046-w.
4. L. Hussain, M. S. Almaraashi, W. Aziz, N. Habib, and S.-U.-R. S. Abbasi, “Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features,” *Waves in Random and Complex Media*, vol. 34, no. 1, pp. 226–251, Mar. 2021, doi: 10.1080/17455030.2021.1905912.
5. A. A.-R. Hussien and N. A. Z. Abdullah, “A review for arabic sentiment analysis using deep learning,” *Iraqi Journal of Science*, vol. 64, no. 12, Dec. 2023, doi: 10.24996/ijs.2023.64.12.37.
6. H. M. Rai, “Cancer detection and segmentation using machine learning and deep learning techniques: A review,” “*Multimedia Tools and Applications*,” vol. 83, no. 9, pp. 27001–27035, Aug. 2023, doi: 10.1007/s11042-023-16520-5.
7. A. Atmakuru *et al.*, “Deep learning in radiology for lung cancer diagnostics: A systematic review of classification, segmentation, and predictive modeling techniques,” *Expert Systems with Applications*, vol. 255, Art. no. 124665, Dec. 2024, doi: 10.1016/j.eswa.2024.124665.
8. P. Hofman *et al.*, “Current challenges and practical aspects of molecular pathology for non-small cell lung cancers,” *Virchows Archiv*, vol. 484, no. 2, pp. 233–246, Oct. 2023, doi: 10.1007/s00428-023-03651-1.
9. N. Kewalramani, C. Machahua, V. Poletti, J. Cadranel, A. U. Wells, and M. Funke-Chambour, “Lung cancer in patients with fibrosing interstitial lung diseases: an overview of current knowledge and challenges,” *ERJ Open Research*, vol. 8, no. 2, Art. no. 50115–52022, Jun. 2022, doi: 10.1183/23120541.00115-2022.
10. X. Feng, Y. Shi, Y. Zhang, F. Lei, R. Ren, and X. Tang, “Opportunities and challenges for inhalable nanomedicine formulations in respiratory diseases: A review,” *International Journal of Nanomedicine*, vol. 19, pp. 1509–1538, Feb. 2024, doi: 10.2147/IJN.S446919.
11. L. L. Dwyer, P. Vadagam, J. Vanderpoel, C. Cohen, B. Lewing, and J. Tkacz, “Disparities in lung cancer: a targeted literature review examining lung cancer screening, diagnosis, treatment, and survival outcomes in the United States,” *Journal of Racial and Ethnic Health Disparities*, vol. 11, no. 3, pp. 1489–1500, May 2023, doi: 10.1007/s40615-023-01625-2.
12. V. K. Gugulothu and S. Balaji, “An early prediction and classification of lung nodule diagnosis on CT images based on hybrid deep learning techniques,” “*Multimedia Tools and Applications*,” vol. 83, no. 1, pp. 1041–1061, May 2023, doi: 10.1007/s11042-023-15802-2.
13. I. Haq *et al.*, “Lung nodules localization and report analysis from computerized tomography (CT) scan using a novel machine learning approach,” *Applied Sciences*, vol. 12, no. 24, Art. no. 12614, Dec. 2022, doi: 10.3390/app122412614.
14. I. Shafi *et al.*, “An effective method for lung cancer diagnosis from ct scan using deep learning-based support vector network,” *Cancers*, vol. 14, no. 21, Art. no. 5457, Nov. 2022, doi: 10.3390/cancers14215457.
15. I. Naseer, S. Akram, T. Masood, A. Jaffar, M. A. Khan, and A. Mosavi, “Performance analysis of state-of-the-art cnn architectures for luna16,” *Sensors*, vol. 22, no. 12, Art. no. 4426, June 2022, doi: 10.3390/s22124426.
16. A. A. Shah, H. A. M. Malik, A. Muhammad, A. Alourani, and Z. A. Butt, “Deep learning ensemble 2D CNN approach towards the detection of lung cancer,” *Scientific Reports*, vol. 13, no. 1, Art. no. 2987, Feb. 2023, doi: 10.1038/s41598-023-29656-z.
17. N. S. Reddy and V. Khanaa, “Intelligent deep learning algorithm for lung cancer detection and classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1747–1754, June 2023, doi: 10.11591/eei.v12i3.4579.
18. I. Naseer, T. Masood, S. Akram, A. Jaffar, M. Rashid, and M. A. Iqbal, “Lung cancer detection using modified AlexNet architecture and support vector machine,” *Computers, Materials & Continua*, vol. 74, no. 1, pp. 2039–2054, Sep. 2022, doi: 10.32604/cmc.2023.032927.
19. L. J. Crasta, R. Neema, and A. R. Pais, “A novel deep learning architecture for lung cancer detection and diagnosis from computed tomography image analysis,” *Healthcare Analytics*, vol. 5, Art. no. 100316, June 2024, doi: 10.1016/j.health.2024.100316.
20. L. Goel and S. Mishra, “A hybrid of modified YOLOv3 with BBO/EE optimizer for lung cancer detection,” “*Multimedia Tools and Applications*,” vol. 83, no. 17, pp. 52219–52251, Nov. 2023, doi: 10.1007/s11042-023-17454-8.

21. E. Shrenitha, M. M. Reddy, D. Sushanth, and M. A. Jabbar, "lung cancer prediction using recurrent residual convolutional neural network and long short term memory," in Proc. IEEE *2024 Int. Conf. on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India, pp. 1–6.
22. J. Chen, S. Liu, and Y. Liu, "ALKU-Net: Adaptive large kernel attention convolution network for lung nodule segmentation," *Electronics*, vol. 13, no. 16, Art. no. 3121, Aug. 2024, doi: 10.3390/electronics13163121.
23. S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions," *Journal of imaging*, vol. 6, no. 12, Art. no. 131, Dec. 2020, doi: 10.3390/jimaging6120131.
24. J. Wang *et al.*, "Preparing CT imaging datasets for deep learning in lung nodule analysis: Insights from four well-known datasets," *Heliyon*, vol. 9, no. 6, Art. no. e17104, June 2023, doi: 10.1016/j.heliyon.2023.e17104.
25. A. Chon, N. Balachandar, and P. Lu, "Deep convolutional neural networks for lung cancer detection," Standford University, pp. 1–9, 2017.
26. H. Jin, C. Yu, Z. Gong, R. Zheng, Y. Zhao, and Q. Fu, "Machine learning techniques for pulmonary nodule computer-aided diagnosis using CT images: A systematic review," *Biomedical Signal Processing and Control*, vol. 79, Art. no. 104104, Jan. 2023, doi: 10.1016/j.bspc.2022.104104.
27. N. Kheradmandi and V. Mehranfar, "A critical review and comparative study on image segmentation-based techniques for pavement crack detection," *Construction and Building Materials*, vol. 321, Art. no. 126162, Feb. 2022, doi: 10.1016/j.conbuildmat.2021.126162.
28. R. Archana and P. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artificial Intelligence Review*, vol. 57, no. 1, Art. no. 11, Jan. 2024, doi: 10.1007/s10462-023-10631-z.
29. S. Ahmed and S. Islam, "Methods in detection of median filtering in digital images: a survey," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 43945–43965, April 2023, doi: 10.1007/s11042-023-14835-x.
30. S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, Art. no. 102036, Aug. 2020, doi: 10.1016/j.bspc.2020.102036.
31. D. N. Chukka, J. S. Meka, S. P. Setty, and P. B. Choppala, "Bayesian selective median filtering for reduction of impulse noise in digital color images," *International Journal of Image and Graphics*, vol. 24, no. 3, Art. no. 2450026, May 2024, doi: 10.1142/S0219467824500268.
32. L. Dénes-Fazakas, S. Csaholczi, G. Eigner, L. Kovács, and L. Szilágyi, "Using resizing layer in u-net to improve memory efficiency," in *Proc. 19thInt. Conf. on Dependability of Computer Systems*, Brunów, Poland, pp. 38–48, July 2024, doi: 10.1007/978-3-031-61857-4_4.
33. L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10173–10196, Aug. 2023, doi: 10.1109/TPAMI.2023.3250241.
34. D. Jaipriya and K. C. Sriharipriya, "Brain computer interface-based signal processing techniques for feature extraction and classification of motor imagery using EEG: A literature review," *Biomedical Materials & Devices*, vol. 2, no. 2, pp. 601–613, May 2023., doi: 10.1007/s44174-023-00082-z.
35. Z. Mahdi, N. Abdalhussien, N. Mahmood, and R. Zaki, "Detection of real-time distributed denial-of-service (DDoS) attacks on internet of things (IoT) networks using machine learning algorithms," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 2139–2159, Aug. 2024, doi: 10.32604/cmc.2024.053542.
36. S. Hu, J. Li, H. Fan, S. Lan, and Z. Pan, "Scale and pattern adaptive local binary pattern for texture classification," *Expert Systems with Applications*, vol. 240, Art. no. 122403, April 2024, doi: 10.1016/j.eswa.2023.122403.
37. A. M. Ayalew, A. O. Salau, B. T. Abeje, and B. Enyew, "Detection and classification of COVID-19 disease from X-ray images using convolutional neural networks and histogram of oriented gradients," *Biomedical Signal Processing and Control*, vol. 74, Art. no. 103530, April 2022, doi: 10.1016/j.bspc.2022.103530.
38. M. M. Bouchene, "Bayesian optimization of histogram of oriented gradients (HOG) parameters for facial recognition," *The Journal of Supercomputing*, vol. 80, pp. 20118–20149, MAy 2024, doi: 10.1007/s11227-024-06259-7.
39. A. Kleppe, O.-J. Skrede, S. D. Raedt, K. Liestøl, D. J. Kerr, and H. E. Danielsen, "Designing deep learning studies in cancer diagnostics," *Nature Reviews Cancer*, vol. 21, no. 3, pp. 199–211, Mar. 2021, doi: 10.1038/s41568-020-00327-9.
40. Z. Y. Khan and Z. Niu, "CNN with depthwise separable convolutions and combined kernels for rating prediction," *Expert Systems with Applications*, vol. 170, Art. no. 114528, May 2021, doi: 10.1016/j.eswa.2020.114528.
41. Q. Zhang, J. Xiao, C. Tian, J. C.-W. Lin, and S. Zhang, "A robust deformed convolutional neural network (CNN) for image denoising," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 331–342, June 2022, doi: 10.1049/cit2.12110.

42. F. A. Mohammed, K. K. Tune, B. G. Assefa, M. Jett, and S. Muhie, "Medical image classifications using convolutional neural networks: a survey of current methods and statistical modeling of the literature," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 699–735, Mar. 2024, doi: 10.3390/make6010033.
43. H. A. A. Mohammed, S. T. Ahmed, R. M. H. Zaki, and Q. K. Kadhim, "Detection and localization of wrist fractures in x-ray imagery using deep learning teaching," *Review of Computer Engineering Research*, vol. 11, no. 3, pp. 85–98, 2024, doi: 10.18488/76.v11i3.3850.
44. Z. K. Abdalrdha, A. M. Al-Bakry, and A. K. Farhan, "A hybrid CNN-LSTM and XGBoost approach for crime detection in tweets using an intelligent dictionary," *Revue d'Intelligence Artificielle*, vol. 37, no. 6, pp. 1651–1661, Dec. 2023, doi: 10.18280/ria.370630.
45. S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 5, Art. no. 102068, June 2024, doi: 10.1016/j.jksuci.2024.102068.
46. Z. S. Mahdi, R. M. Zaki, and L. Alzubaidi, "Advanced hybrid techniques for cyberattack detection and defense in IoT networks," *Security and Privacy*, pp. 1–19, Oct. 2024, doi: 10.1002/spy2.471.
47. R. Hoque, S. Das, M. Hoque, and E. Haque, "Breast cancer classification using XGBoost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985–1994, 2024, doi: 10.30574/wjarr.2024.21.2.0625.
48. R. M. Zaki and I. S. Naser, "Hybrid classifier for detecting zero-day attacks on IoT networks," *Mesopotamian Journal of CyberSecurity*, vol. 4, no. 3, pp. 59–74, Nov. 2024, doi: 10.58496/MJCS/2024/016.
49. W. Jiao, X. Hao, and C. Qin, "The image classification method with CNN-XGBoost model based on adaptive particle swarm optimization," *Information*, vol. 12, no. 4, Art. no. 156, April 2021, doi: 10.3390/info12040156.
50. K. Suriyan and R. Nagarajan, "Particle swarm optimization in biomedical technologies: innovations, challenges, and opportunities," *Emerging Technologies for Health Literacy and Medical Practice*, pp. 220–238, Feb. 2024, doi: 10.4018/979-8-3693-1214-8.ch011.