1

Imaging and Radiation Research

EnPress

https://systems.enpress-publisher.com/index.php/IRR

2024 Volume 7 Issue 1
ISSN: 2578-1618 (Online)

Editorial Board

Editor-in-Chief

Associate Editors-in-Chief

Editorial Board Members

Volume 7 Issue 1·2024

Imaging and Radiation Research

EnPress Publisher

EnPress Publisher

2024 | Volume 7 | Issue 1

Imaging and Radiation Research

https://systems.enpress-publisher.com/index.php/IRR

Contents

Articles

Perspective

Imaging and Radiation Research 2024, 7(1), 6195.
https://doi.org/10.24294/irr6195

Article

Enhancing breast cancer detection in thermographic images using deep hybrid networks

Rezazadeh Hanieh, Saniei Elham*, Salehi Barough Mehdi

Department of Nuclear Engineering, Central Tehran Branch, Islamic Azad University, Tehran 14174, Iran
*** Corresponding author:** Saniei Elham, el.saniei@iau.ac.ir, elhsaniei@gmail.com

CITATION

Hanieh R, Elham S, Mehdi SB. (2024). Enhancing breast cancer detection in thermographic images using deep hybrid networks. Imaging and Radiation Research. 7(1): 6195. https://doi.org/10.24294/irr6195

ARTICLE INFO

Received: 6 March 2024
Accepted: 20 April 2024
Available online: 4 May 2024

Abstract: Breast cancer was a prevalent form of cancer worldwide. Thermography, a method for diagnosing breast cancer, involves recording the thermal patterns of the breast. This article explores the use of a convolutional neural network (CNN) algorithm to extract features from a dataset of thermographic images. Initially, the CNN network was used to extract a feature vector from the images. Subsequently, machine learning techniques can be used for image classification. This study utilizes four classification methods, namely Fully connected neural network (FCnet), support vector machine (SVM), classification linear model (CLINEAR), and KNN, to classify breast cancer from thermographic images. The accuracy rates achieved by the FCnet, SVM, CLINEAR, and k-nearest neighbors (KNN) algorithms were 94.2%, 95.0%, 95.0%, and 94.1%, respectively. Furthermore, the reliability parameters for these classifiers were computed as 92.1%, 97.5%, 96.5%, and 91.2%, while their respective sensitivities were calculated as 95.5%, 94.1%, 90.4%, and 93.2%. These findings can assist experts in developing an expert system for breast cancer diagnosis.

Keywords: breast cancer detection; deep learning; hybrid network; thermography images; convolutional neural network

1. Introduction

Breast cancer was one of the most prevalent forms of cancer worldwide. Early detection plays a crucial role in successful treatment. Thermography was a method for imaging breast cancer. It employs an infrared camera to capture temperature patterns in the target area. This technique was both safe and cost-effective compared to other imaging methods. However, it had limitations such as a relatively high rate of false positives and false negatives (around 10%), making accurate determination of affected areas challenging [1]. Recent advancements in this field include the detection of areas with high temperature gradients, automated identification of desired areas within each breast, and analysis of asymmetry [2].

Several deep learning techniques had been proposed for accurate breast cancer diagnosis, including multi-layer perceptron neural networks, convolutional neural networks, and fuzzy neural network expert systems [3]. These techniques had been evaluated using diverse datasets and features, such as histopathological images, mammography images, and thermograms. Moreover, researchers had explored the integration of artificial intelligence-based tools in clinical practice to enhance the accuracy and efficiency of breast cancer screening and grading. Desai and Shah introduced a novel approach for breast cancer diagnosis that employs deep learning techniques [4]. Their study used the efficacy of multi-layer perceptron (MLP) and convolutional neural network (CNN) models in classifying mammography into benign

and malignant classes. The findings revealed that CNN outperformed MLP in terms of accuracy for cancer detection. Algehine and colleagues introduced a fuzzy neural network expert system for early detection of breast cancer in mammography [5]. Their approach combined fuzzy logic, neural networks, and machine learning algorithms to achieve high accuracy in diagnosing early-stage breast cancer. In summary, these studies highlight the potential of deep learning techniques, such as MLP, CNN, and fuzzy neural network expert systems, for accurate breast cancer diagnosis. The integration of artificial intelligence-based tools in clinical practice had showed promise, but further investigation was required to ensure safe and effective implementation.

Convolutional neural networks (CNNs) had played a crucial role in establishing non-linear mappings between input and output, autonomously learning local and high-level features through multilayer network architectures, as well as predefined feature sets. In a study [6], a deep learning-based approach utilizing CNNs was proposed for early breast cancer detection, achieving a remarkable classification accuracy in distinguishing between benign and malignant classes. Riggio et al. explored the current understanding of metastatic breast cancer and addressed unresolved challenges that need to be tackled to improve patient outcomes. Despite the complexity and computational slowness caused by simultaneous use of different algorithms, their study achieved an accuracy of 98% in accurately differentiating cancerous parts from healthy breast tissue [7].

Gonçalves and others used pretrained convolutional neural networks such as VGG16, Densenet201, and ResNet50 to classify thermography images. The DenseNet model did the best, with an accuracy of 91.67%, sensitivity of 100%, and specificity of 83.3%. This study showed that using deep learning models was effective for detecting breast cancer. They used 38 pictures for each category [8]. Shahnaz et al. reviewed naive bayes, SVM, logistic regression, KNN, random forest neural networks, MLP and CNN classifiers for the detection of breast cancer [9]. The CNN had the highest accuracy at 98.06%, while the accuracy of MPL was 97.891 at five layers. This showed that CNN was better than other methods. Desai et al. used MLP and CNN for mammography image classification, achieving an accuracy of 93.6%, with CNN outperforming MLP [4]. In order to increase diagnostic accuracy, research incorporated machine learning and deep learning algorithms. They used the effectiveness of their approach in diagnosing metastasis using various features extracted from histopathological images, including color, texture, and morphology. The results exhibited high diagnostic accuracy (96.8%), highlighting the potential benefits of hybrid approaches.

Dey et al. made a model called DenseNet121 and added two detectors (Prewitt and Roberts) to convert input from thermal images. It does really well with 98.8% accuracy on the DMR-IR dataset, doing better than other ways people had tried [10]. In a study introduced a new computer Aided System that used deep learning to find breast cancer [10]. It used all breast thermogram views and patient information. In this research, they used AlexNet to analyze thermograms, and a classic neural network for clinical data. The findings showed that using more than one input did better than using just one input, and the overall accuracy was 90.48%, with a sensitivity of 93.33%. The approach to doing something had changed compared to past references. While recent

models [10] show better accuracy, it may not be suitable for devices with limited memory due to its numerous parameters. Also, these models were time consuming on a high-dimensional data set that includes many predictor variables. Our recommendation for resolving these issues was to employ the deep hybrid network. To identify important features, we designed a CNN and then used some machine learning (ML) techniques to categorize the patterns into separate groups. The use of the CNN allows for the extraction of significant features by producing detailed features that can be used in combination with ML classifiers and undergo thorough testing. The essential aspect of ML classifiers was that they were fast. This paper had made important contributions.

A simple CNN model was made with limited parameters, instead of pre-trained network that can be used on a mobile device.

Four deep learning-based identification method of breast cancer from thermal images was proposed by combining CNN and different ML classifiers: Fully connected neural network (FCnet), support vector machine (SVM), classification linear model (CLINEAR), k-nearest neighbor (KNN)

A comparison was made between the proposed models and other related works.

2. Methodology

This research proposed a hybrid strategy to recognize breast cancer from thermographic Images. The recommended procedure was illustrated in **Figure 1**. Deep learning was a type of machine learning that takes inspiration from the way the brain works. Convolutional neural networks (CNN) were the most essential types of deep neural networks designed to process and predict various features simultaneously. They had used remarkable capabilities in extracting meaningful features from images. we proposed the utilization of CNNs for feature extraction from thermographic images. Then the features were classified using four different ML classifiers for choosing the best hybrid method. The suggested methodology was divided into five steps were talked about further down.

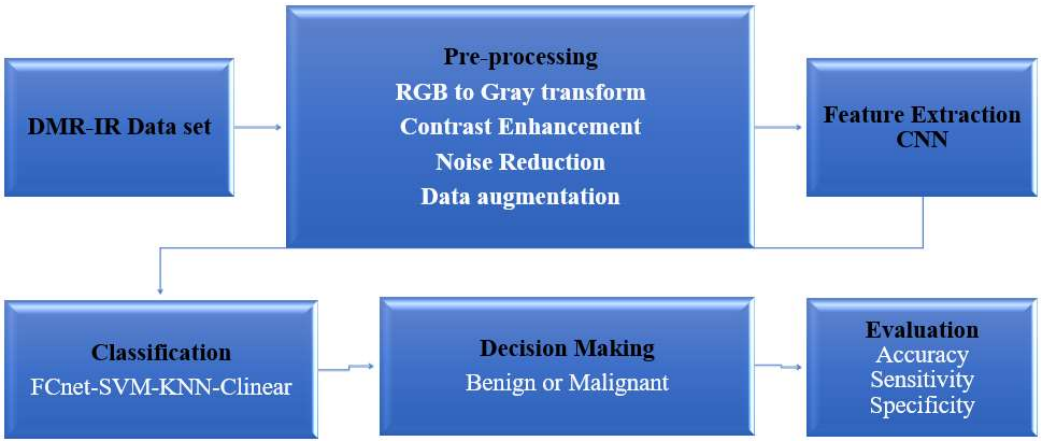


Figure 1. The illustration of the proposed method.

2.1. Pre-processing

Preparing the images before using it in machine learning was important, and preprocessing was the first step in this process. Several preprocessing methods were used in this study, which were mentioned below.

- Grayscaling: Images data can be simplified and computational requirements reduced by converting color images to grayscale. The images from the dataset were colored. Therefore, they were converted to gray images at first in **Figure 2b**.
- Contrast enhancement: After grayscaling, the contrast of the images was adjusted using histogram equalization as shown in **Figure 2c**.
- Noise reduction: Median filter was applied to remove unwanted noise from the images. It analyzes the image pixel by pixel, and replaces each pixel with the median of neighboring entries. The smoothed image was shown in **Figure 2d**.
- Data augmentation: In machine learning, "imbalanced classes" was a familiar problem particularly occurring in classification. Ideally, all classes would had an equal number of observations. However, the classes in data set were imbalanced (2800 cancer images, 4500 normal images) and if not handled correctly, this imbalance can be detrimental to the learning process because the learning was biased in favor of the dominant classes. To handle this issue, data augmentation technique was employed. Data augmentation was a way to increase the number of training images by manipulating the original image. In this study, this involved scaling up the original image size by 50%, applying random rotations of up to 20 degrees in any direction, and introducing random translations of up to a maximum of 3 pixels. This was applied on the images from the cancer classes.

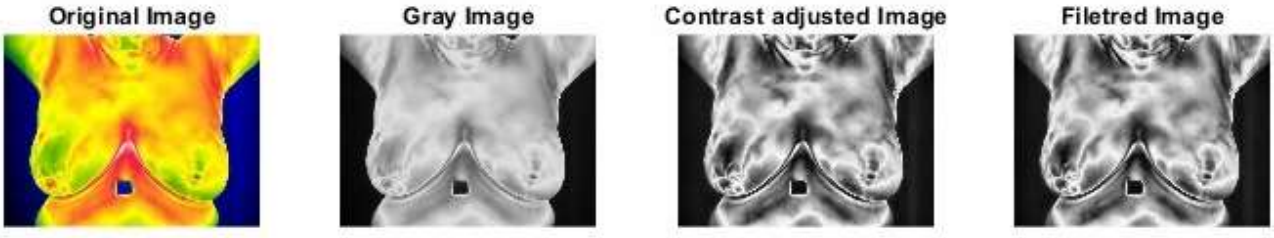


Figure 2. A example of pre-processing steps: **(a)** original image; **(b)** gray image; **(c)** contrast adjusted image; **(d)** median filtered image.

2.2. CNN proposed architecture for feature extraction

CNNs had used remarkable capabilities in extracting meaningful features from images. They were one of the most essential types of deep neural networks designed to process and predict various features simultaneously. In this research, a CNN architecture with four convolutional layers was employed. It consists of the normalization, pooling and two fully connected layers. **Figure 3** provides an overview of the network layout. It will be described as follows:

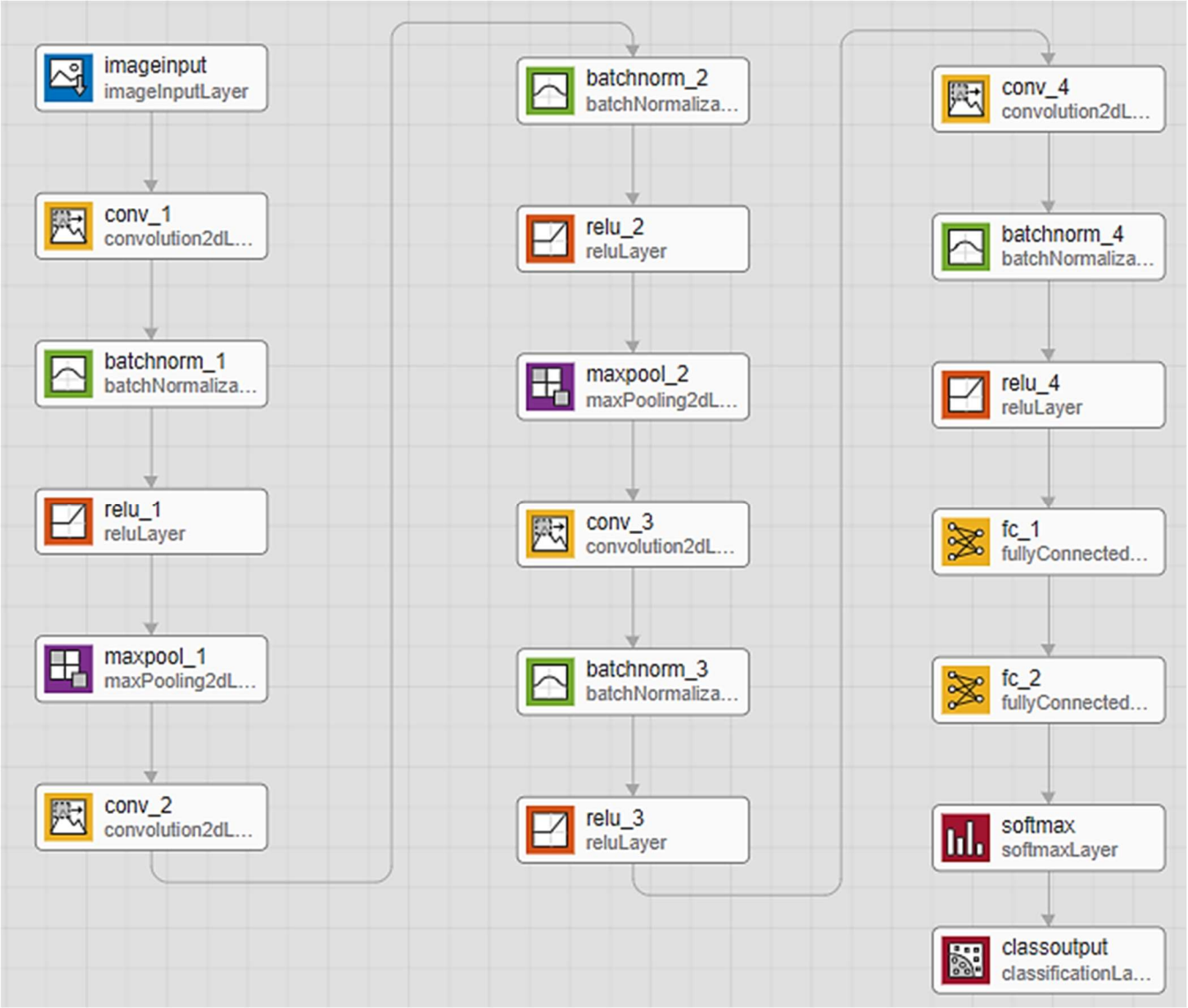


Figure 3. The diagram of the proposed convolutional neural network.

1) The first convolutional layer incorporates 8 filters with a kernel size of 77 pixels. Additionally, we used the padding option, which expands the border pixels before the convolution operation.
2) Stacked convolutional layers were accompanied by a batch normalization operation. After each convolutional layer, a modified linear unit operation was applied. The network also includes a max-pooling layer with a kernel size of 22 pixels and a stride of 2.
3) The subsequent stacked convolutional layers and the batch normalization layer follow the same pattern as the first layer. However, the convolution kernel sizes were set to 55, 33, and 33, respectively. The filter size and number for each convolutional layer was given in **Table 1**.

Table 1. The layers of the proposed CNN and their parameters.

Name	Type	Activation shape	Learnable parameters
Image input	Image Input	480 × 640 × 3	-
conv1	Convolution	480 × 640 × 8	Weights 7 × 7 × 3 × 8 Bias 1 × 1 × 8
batchnorm_1	Batch Normalization	480 × 640 × 8	Offset 1 × 1 × 8 Scale 1 × 1 × 8
relu_1	ReLU	480 × 640 × 8	-
maxpool_1	Max Pooling	240 × 320 × 8	-
conv2	Convolution	240 × 320 × 8	Weights 7 × 7 × 8 × 8 Bias 1 × 1 × 8
batchnorm_2	Batch Normalization	240 × 320 × 8	Offset 1 × 1 × 8 Scale 1 × 1 × 8
relu_2	ReLU	240 × 320 × 8	-
maxpool_2	Max Pooling	120 × 160 × 8	-
conv3	Convolution	120 × 160 × 8	Weights 7 × 7 × 8 × 8 Bias 1 × 1 × 8
batchnorm_3	Batch Normalization	120 × 160 × 8	Offset 1 × 1 × 8 Scale 1 × 1 × 8
relu_3	ReLU	120 × 160 × 8	-
conv4	Convolution	120 × 160 × 8	Weights 7 × 7 × 8 × 8 Bias 1 × 1 × 8
batchnorm_4	Batch Normalization	120 × 160 × 8	Offset 1 × 1 × 8 Scale 1 × 1 × 8
relu_4	ReLU	120 × 160 × 8	-
fc_1	Fully Connected	1 × 1 × 16	Weights 16 × 1153600 Bias 1 × 1 × 16
fc_2	Fully Connected	1 × 1 × 2	Weights 2 × 16 Bias 1 × 1 × 2
SoftMax	SoftMax	1 × 1 × 2	-
Class output	Classification Output	-	-

Training hyperparameters:

The CNN network was trained for 100 epochs. In order to stabilize the network during the initial training phase, a low learning rate (0.01) was used initially, gradually increasing over time. The ‘adam’ optimizer was selected over the ‘sgdm’ optimizer because it’s a combination of two different optimizers, rmsprop and adagrad. To train and test the CNN network, an 80:20 split was employed, with 80% of the dataset allocated for training and 20% for testing. This ratio was commonly used in machine learning programs. Additionally, to mitigate the risk of overfitting, a cross-fold validation method with a ratio of 5 was employed. There was a list of hyperparameters in **Table 2**.

Table 2. Training hyper parameters.

Hyper parameters	Specifications
Epoch	100
Initial learning rate	0.01
MiniBatchSize	64
Optimizer	Adaptive moment estimation (Adam)
Validation frequency	10

2.3. Classifiers

After applying the CNN on each image, the corresponding feature vector was obtained. In this research, to compare the accuracy and speed of different classifiers, four machine learning methods were employed: fully connected neural network (FCnet), support vector machine (SVM), classification linear model (CLINEAR), k-nearest neighbor (KNN). Typically, at the end of the CNN, a fully connected neural network (FCnet) was used to classify the images. This approach balances speed and accuracy, aligning with the characteristics of the convolutional network. The first fully connected layer consists of 16 neurons and the second fully connected layer contains 2 neurons. The output layer employs a SoftMax activation function, with output values representing the probabilities of belonging to each group (0 or 1) which effectively segregate the images into normal and cancer groups.

Support vector machine (SVM) was a supervised learning method commonly used for classification. It had used superior performance compared to older classification methods in recent years. The SVM classifier operates based on linear classification of data. Various kernel functions, such as exponential, polynomial, and sigmoid kernels, can be used to generate these boundaries, thereby increasing the complexity and accuracy of the SVM method.

The k-nearest neighbor (KNN) algorithm was a non-parametric statistical method commonly used for statistical classification and regression. This algorithm selects k closest training examples in the data space, and its output varies depending on the type used for classification or regression. In the classification mode, given a specified value for k, it calculates the distance between the point we want to label and its closest neighbors. Based on the maximum number of votes from these neighboring points, the algorithm makes a decision regarding the label of the point. Euclidean distance was typically used to calculate this distance.

A classification linear classifier (CLINEAR) was a model that categorizes a set of data points into discrete classes based on a linear combination of their variables. This method minimizes the objective function using techniques that reduce computation time, such as stochastic gradient descent. **Table 3** provides information on the number of features associated with each classifier.

Table 3. Classifier parameters in the proposed method.

Classifier	Function	Features dimension
FCnet	SoftMax Layer	2 × 4344
SVM	fitcSVM	2 × 4344
CLINEAR	fitclinear	2 × 4344
KNN	fitcknn	2 × 4344

Overall, these different classifiers provide varying approaches to image classification, each with its advantages and considerations. By comparing their performance, we can gain insights into their accuracy and speed for the given task.

3. Results analysis

(1) Dataset

The study was based on the DMR-IR database [11], which was obtained from volunteers in Brazil by the Federal University of Fluminense. For research purposes, the dataset was publicly available and its collection was ethically approved. These images originated from diverse sources, such as hospitals, clinics, and research institutes, and encompassed a wide range of age and gender groups. Patient-related information, including age and gender, was available for most images, which provides valuable data for the development of breast cancer detection algorithms. The dataset provided by this group was widely recognized for its accuracy and reliability, making it a valuable resource for academic and professional research. The thermographic images in the database were captured using a FLIR SC-620 camera with a resolution of 480 × 640 and a thermal sensitivity of 40 mK. During the image processing stage, all images were converted to grayscale. For each individual, images were taken from angles of 45°, 90° to the right, and 90° to the left, resulting in a total of five thermographic images per person (**Figure 4**). The dataset used in this study comprised thermography images of 4500 healthy individuals and 2800 individuals diagnosed with cancer.

(a) **(b)** **(c)** **(d)** **(e)**

Figure 4. Example of thermographic images utilized: **(a)** front view; **(b)** right 45-degree angle; **(c)** right 90-degree angle; **(d)** left 45-degree angle; **(e)** left 90-degree angle.

(2) Performance evaluation

The research algorithms were implemented using MATLAB 2021 programming language. **Figure 5** illustrates the accuracy and error graphs at each stage of training. It was evident that increasing the training steps leads to a reduction in losses and higher accuracy. Additionally, **Table 4** presents a comparison of the speed and accuracy of the results for each classifier. The results indicate that the KNN classifier was approximately twice as fast as the concrete CNN. The experimental setup used a Windows system with 8 GB RAM, an Intel(R) Core i5-4430 CPU@3.00GHz x64-based processor. Despite the CNN network demonstrating good accuracy, it exhibits slower speed compared to other methods.

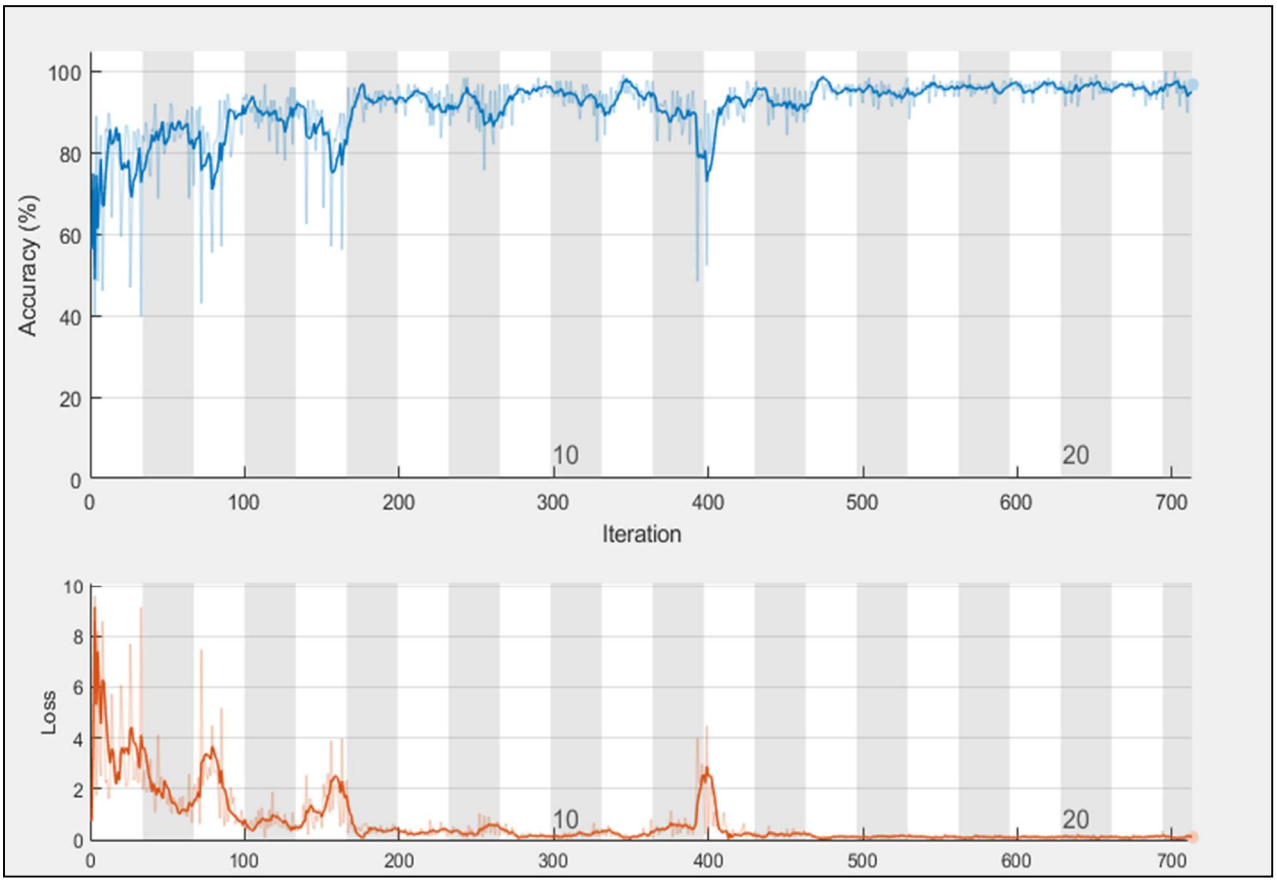
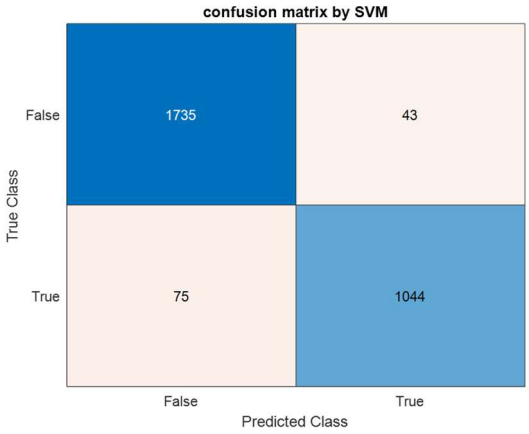


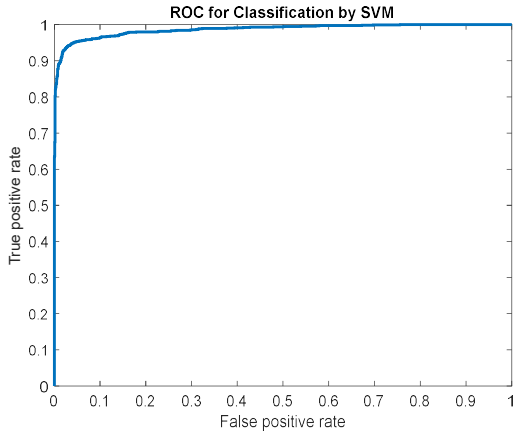
Figure 5. Training progress graph of the proposed CNN.

Table 4. Comparison of learning speed in cancer diagnosis using different classification methods.

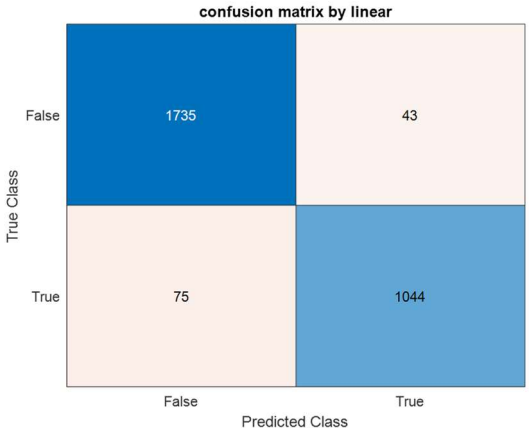
Classifier	Speed (second)
FCnet	336.1
SVM	205.7
CLINEAR	165.5
KNN	162.7

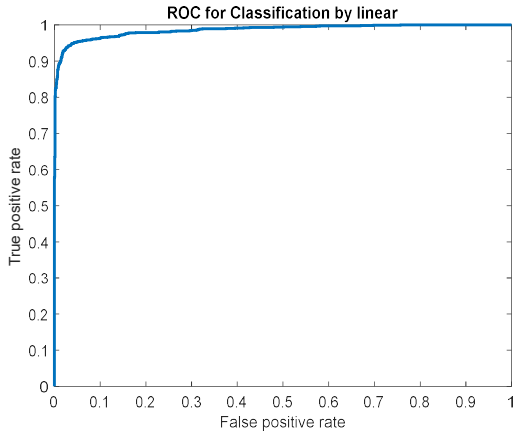
The performance of all classifiers was evaluated using ROC curves and confusion matrices (**Figure 6**). The SVM and CLINEAR methods yielded nearly identical results, with a total of 118 misclassified individuals. However, the KNN classifier resulted in a higher misclassification rate, with 140 misclassified individuals. **Table 5** provides further insights into the accuracy of the testing process. It shows that the SVM and CLINEAR classifiers exhibited higher training accuracy compared to other classifiers.



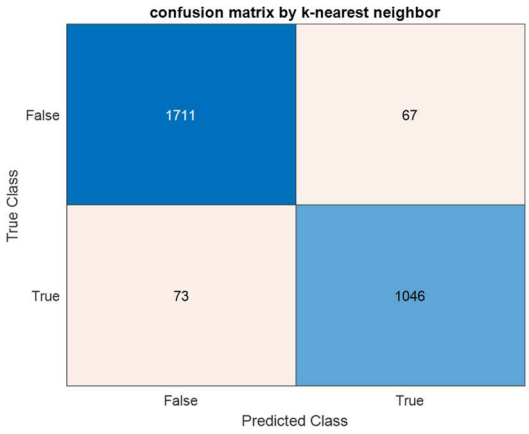


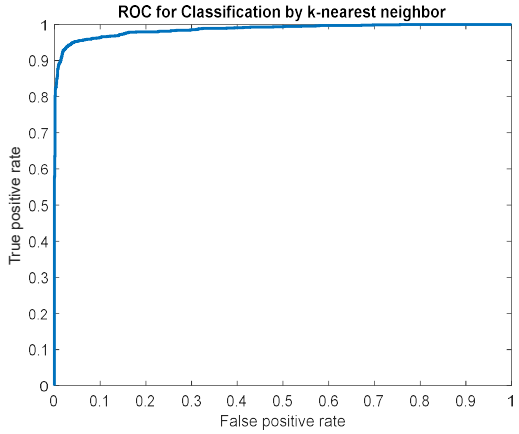
(a)





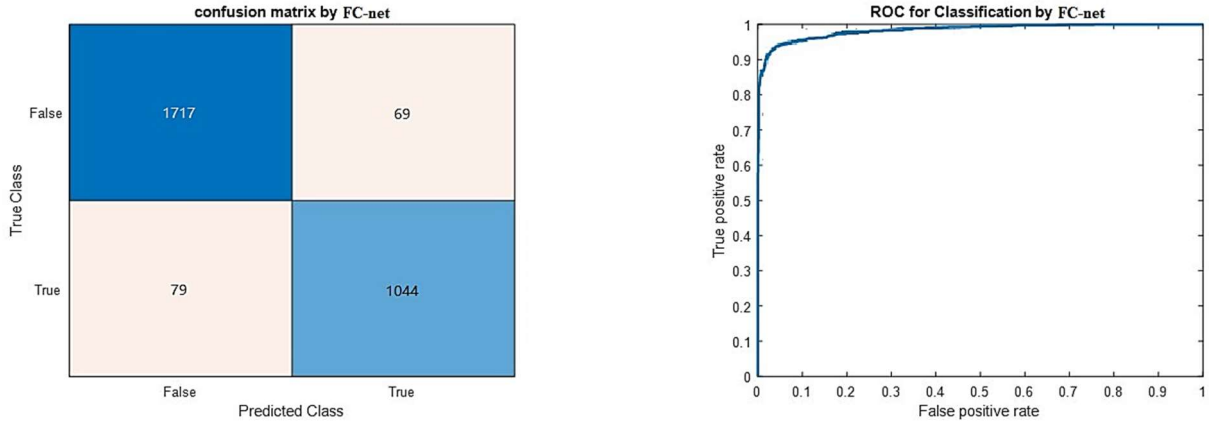
(b)





(c)

Figure 6. (*Continued*).



(d)

Figure 6. ROC curve and confusion matrix for image classification using. **(a)** SVM; **(b)** CLINEAR; **(c)** KNN; and **(d)** FC-Net.

Table 5. Performance comparison of different hybrid methods.

Hybrid method	Accuracy	Sensitivity	Specificity
CNN-FCnet	94.2%	93.2%	91.2%
CNN-SVM	95.0%	90.4%	96.5%
CNN-CLINEAR	95.0%	94.1%	97.5%
CNN-KNN	94.1%	95.5%	92.1%

4. Conclusion

Several papers [4,5,10,12,13] experimented on the thermal images to detect breast cancer. **Table 6** compares the current methods of deep learning for breast cancer detection. Tsietso et al. [12] use a variety of deep learning techniques for cancer detection from thermographic images. Transfer learning was used by Dey et al. [10] and feature extraction was done by pretrained VGG16, VGG19. Pre-trained models in transfer learning were complicated and had lots of parameters. The suggested method in some studies [1–6] been shown to be more accurate but cannot be used with memory constrained devices due to its high number of parameters. In contrast to a previously trained network, the proposed method is uncomplicated. Creating a smaller network is good because it can help to use algorithms on mobile devices. It is possible to use an automated algorithm such as CNN for the extraction of features, since it is capable of producing deep learning features that can be used for the evaluation of ML classifiers and a comprehensive evaluation. The essential aspect of ML classifiers is that they are fast.

Table 6. Comparison of the proposed hybrid methods with existing methods.

Study	Year	Methodology	Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
Abdel-Nasser M [13]	2019	CAD (Computer-Aided Diagnostic), ML	DMR-IR Dataset	95.8		94.6
Algehyne EA et al. [5]	2022	Fuzzy NN Expert System	Wisconsin breast cancer database	95.5	93.8	94.9
Dey S et al. [10]	2022	DenseNet121+, VGG16, VGG19	DMR-IR	98.8	98	
Tsietso D et al. [12]	2023	CADx, DNN, AlexNet	DMR-IR Dataset	90.48	93.33	
Desai M and Shah M [4]	2023	MLP, CNN	Kaggle data set (BC)	93.6	92.1	95.4
Awotunde JB et al. [14]	2023	Hybrid ML & DL	Histopathological images	96.8	94.5	96.0
Gonçalves CB et al. [15]	2022	VGG-16, Densenet 201, and Resnet 502	Thermography	91.67	100	83.3
Our propose method	2024	CNN + (SVM, CLINEAR, KNN)	DMR-IR Dataset	94.2–95.0	90.4–95.5	91.2–97.5

The proposed algorithms employed a CNN with 4th layers to detect relevant features from input images. The extracted features were then fed into the four ML classifier for the purpose of breast cancer detection. The results indicate that both SVM and CLINEAR classifiers yield similar outcomes, with a total of 118 misdiagnosed individuals. On the other hand, the KNN method results in 140 misdiagnosed cases. It is worth noting that the training accuracy of SVM and CLINEAR algorithms surpasses that of other networks. Nevertheless, the FCnet classifier also exhibits high accuracy, outperforming the KNN method by a margin of 0.1%. This improved accuracy can be attributed to the object detection kernel used in the convolutional algorithm, which proves particularly effective for high-resolution images. Additionally, it is important to highlight that the network executed the image only once, contributing to the speed of the CNN, especially when running on parallel processing cards, enabling real-time processing. Furthermore, the findings reveal that the KNN method demonstrates higher sensitivity compared to the other methods, whereas the SVM method exhibits the lowest sensitivity. In terms of false positive rates, the FCnet method performs better than all other methods, while the CLINEAR method yields higher rates than the remaining approaches.

The differences in the effects of the various classification methods used in this study can be attributed to their unique algorithmic structures and operational mechanisms. FCnet demonstrates superior accuracy in classifying thermographic images due to their ability to extract hierarchical and meaningful features from raw data through multiple convolutional layers. This capability allows CNNs to capture intricate patterns and variations within the images, making them particularly effective for complex data sets. In contrast, SVM and CLINEAR classifiers operate by identifying optimal hyperplanes that maximize the margin between different classes. This approach was robust for high-dimensional feature spaces and provides reliable performance, although it may not be as adept as FCnet in handling non-linear and highly complex data patterns. KNN, a non-parametric method, classifies data points based on the majority vote of their nearest neighbors, making it simple and effective

for smaller datasets. However, KNN's performance can degrade with larger datasets due to the increased computational cost during the prediction phase. Comparatively, traditional methods like SVM and KNN exhibit strengths in specific scenarios but may fall short in versatility and accuracy when compared to complicated approaches like FCnet. This comparative analysis underscores the importance of selecting the appropriate classifier based on the dataset characteristics and the specific requirements of the diagnostic application.

This comparative summary highlights the competitive performance of our proposed method and the potential of integrating newer deep learning architectures and hybrid models to further enhance breast cancer detection using thermographic images. Future research should focus on leveraging these advancements and validating the approach on larger, more diverse datasets to ensure robust and reliable performance in clinical settings. Additionally, incorporating other classification methods such as genetic algorithms in conjunction with convolutional networks was suggested as a potential avenue for investigation.

Author contributions: Conceptualization, RH and SE; methodology, RH; software, RH; validation, RH, SE and SBM; formal analysis, RH; investigation, RH; resources, SE; data curation, SE; writing—original draft preparation, RH; writing—review and editing, RH; visualization, RH; supervision, SE; project administration, SE; funding acquisition, RH. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Mousavi H, Bagherian R. Health literacy and breast cancer. Health Psychology. 2019; 8(31): 91-102.
2. Mohamed AA, Berg WA, Peng H, et al. A deep learning method for classifying mammographic breast density categories. Medical Physics. 2018; 45(1): 314-321. doi: 10.1002/mp.12683
3. Clady X, Negri P, Milgram M, Poulenard R. Multi-class vehicle type recognition system. In: Proceedings of the Artificial Neural Networks in Pattern Recognition: Third IAPR Workshop, ANNPR 2008; 2-4 July 2008; Paris, France. pp. 228-239.
4. Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth. 2021; 4: 1-11. doi: 10.1016/j.ceh.2020.11.002
5. Algehyne EA, Jibril ML, Algehainy NA, et al. Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. Big Data and Cognitive Computing. 2022; 6(1): 13. doi: 10.3390/bdcc6010013
6. Aidossov N, Zarikas V, Mashekova A, et al. Evaluation of Integrated CNN, Transfer Learning, and BN with Thermography for Breast Cancer Detection. Applied Sciences. 2023; 13(1): 600. doi: 10.3390/app13010600
7. Riggio AI, Varley KE, Welm AL. The lingering mysteries of metastatic recurrence in breast cancer. British Journal of Cancer. 2020; 124(1): 13-26. doi: 10.1038/s41416-020-01161-4
8. Gonçalves CB, Souza JR, Fernandes H. Classification of static infrared images using pre-trained CNN for breast cancer detection. In: Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS); 7 June 2021. pp. 101-106.
9. Shahnaz C, Hossain J, Fattah SA, et al. Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database. In: Proceedings of the 2017 IEEE region 10 humanitarian technology conference (R10-HTC); 21 Decemebr 2017. pp. 792-797.

10. Dey S, Roychoudhury R, Malakar S, et al. Screening of breast cancer from thermogram images by edge detection aided deep transfer learning model. Multimedia Tools and Applications. 2022; 81(7): 9331-9349. doi: 10.1007/s11042-021-11477-9
11. DMI: Visual Computing Group. Available online: https://visual.ic.uff.br (accessed on 28 April 2024).
12. Tsietso D, Yahya A, Samikannu R, et al. Multi-Input Deep Learning Approach for Breast Cancer Screening Using Thermal Infrared Imaging and Clinical Data. IEEE Access. 2023; 11: 52101-52116. doi: 10.1109/access.2023.3280422
13. Abdel-Nasser M, Moreno A, Puig D. Breast Cancer Detection in Thermal Infrared Images Using Representation Learning and Texture Analysis Methods. Electronics. 2019; 8(1): 100. doi: 10.3390/electronics8010100
14. Awotunde JB, Panigrahi R, Khandelwal B, et al. Breast cancer diagnosis based on hybrid rule-based feature selection with deep learning algorithm. Research on Biomedical Engineering. 2023; 39(1): 115-127. doi: 10.1007/s42600-022-00255-7
15. Gonçalves CB, Souza JR, Fernandes H. CNN architecture optimization using bio-inspired algorithms for breast cancer detection in infrared images. Computers in Biology and Medicine. 2022; 142: 105205. doi: 10.1016/j.compbiomed.2021.105205

Imaging and Radiation Research 2024, 7(1), 5700.
https://doi.org/10.24294/irr5700

Article

Collaborative intelligent decision systems for safe and reliable AI-assisted medical image diagnostics

Serge Dolgikh

National Aviation University, 25005 Kropyvnytskyi, Ukraine; serged.7@gmail.com

CITATION

Dolgikh S. (2024). Collaborative intelligent decision systems for safe and reliable AI-assisted medical image diagnostics. Imaging and Radiation Research. 7(1): 5700. https://doi.org/10.24294/irr5700

ARTICLE INFO

Received: 7 April 2024
Accepted: 19 May 2024
Available online: 30 May 2024

Abstract: The cost of diagnostic errors has been high in the developed world economics according to a number of recent studies and continues to rise. Up till now, a common process of performing image diagnostics for a growing number of conditions has been examination by a single human specialist (i.e., single-channel recognition and classification decision system). Such a system has natural limitations of unmitigated error that can be detected only much later in the treatment cycle, as well as resource intensity and poor ability to scale to the rising demand. At the same time Machine Intelligence (ML, AI) systems, specifically those including deep neural network and large visual domain models have made significant progress in the field of general image recognition, in many instances achieving the level of an average human and in a growing number of cases, a human specialist in the effectiveness of image recognition tasks. The objectives of the AI in Medicine (AIM) program were set to leverage the opportunities and advantages of the rapidly evolving Artificial Intelligence technology to achieve real and measurable gains in public healthcare, in quality, access, public confidence and cost efficiency. The proposal for a collaborative AI-human image diagnostics system falls directly into the scope of this program.

Keywords: image diagnostics; machine learning; transfer learning; collaborative Human-AI systems; intelligent decision systems; AIM

1. Introduction

The cost of diagnostic errors has been high in the developed world economics according to a number of recent studies and continues to rise [1]. Up till now, a common process of performing image diagnostics for a growing number of conditions has been examination by a single human specialist (i.e., single-channel recognition and classification decision system). Such a system has natural limitations of unmitigated error that can be detected only much later in the treatment cycle, as well as resource intensity and poor ability to scale to the rising demand [2]. At the same time Machine Intelligence (ML, AI) systems, specifically those including deep neural network and large visual domain models have made significant progress in the field of classification and recognition of complex data, in many instances achieving the level of an average human and in a growing number of cases, a human specialist in the effectiveness of image recognition tasks [3,4].

Machine Intelligence models and systems (ML, AI) have a number of essential strengths, such as: stability and resilience with respect to the environmental factors and influences; superior operating performance in both time and volume; shorter training time and time to operation; accuracy in execution of intelligent tasks approaching and in a growing number of applications surpassing that of a human specialist; cost efficiency in operation. In a number of applications, ML/AI systems

demonstrated the ability to identify characteristic types or "concepts" in complex realistic data [5,6] that can be instrumental in the analysis and description of its information structure.

On the other hand, integration of machine intelligence models in essential and critical public interest applications which include public health care is less than straightforward and can be impeded by insufficient understanding of the processes that lead to their decisions (explainability and trust challenges [7,8]), preparedness of the public and general public trust, dependence on large bodies of trusted data describing the domain in sufficient detail and others.

A promising avenue in harnessing the power and the potential of machine intelligence in public interest applications has been developed over the years in the area of collaborative intelligent systems [9,10]. An inherent promise of this approach is the potential to use the machine and human intelligences in a collaborative process that used the respective strengths of each type while mitigating their downsides. It was shown that by designing decision systems in such a way that both human and machine components could contribute to the success of the resulting decision, a synergetic effect can be achieved with significant improvement in the accuracy, and as a consequence, a noticeable reduction in the diagnostic error and the associated with it cost [11].

The objectives of the AI in Medicine (AIM) program [12] were set to leverage the opportunities and advantages of the rapidly evolving Artificial Intelligence technology to achieve real and measurable gains in public healthcare, in quality, access, public confidence and cost efficiency. The proposal falls directly into the scope of this program.

2. Drivers of AI integration in public health care

The justification for the research in collaborative AI-human decision systems, as was briefly mentioned previously, is based on four drivers creating an opportunity for a successful integration of Machine Intelligence technology in the tasks of image diagnostics in the public healthcare system, with the potential to improve, measurably and significantly, the accuracy and productivity of the diagnostics tasks and processes.

Cost and Resources: rising cost of diagnostic errors and constraints on available resources in public healthcare systems.

Technology: recent advances in Machine Intelligence technology, bringing the level of image recognition success to that of a human.

Complementarity: complementary, mutually contributing and complementing nature of human and machine intelligences creates an opportunity for cooperative and collaborative work process with improved outcomes in the targeted tasks.

Trust and safety: the solution has to have uncompromised safety and full human control over the resulting decision.

The drivers and opportunities associated with them set the foundation for a collaborative, joint decision framework of human and machine intelligences that can use the strengths of either while mutually compensating for the limitations and shortcomings, to improve the outcome of the collective decision in image diagnostic tasks. The potential and the window of opportunity for the development and

integration of collaborative AI-human intelligent decision systems is illustrated in **Figure 1**.

Figure 1. Drivers for the advent of synergetic collaborative AI-human decision systems.

The drivers described above lay the ground and provide the incentive for research into intelligent systems that could harness the strong sides of either type of intelligence to produce superior outcome, within the framework of expectations and requirements, to the use of either system on its own.

3. Complementarity and synergetic potential of collaborative intelligent systems

A collaborative intelligent decision system outlined in this article is based on the observation that due to the complementary nature of the human and machine intelligences they may not be expected to make "many" mistakes in the same situations and cases; and as a consequence, an opportunity emerges for the creation of synergetic intelligent systems where human and machine channels would be able to complement and correct each other.

For example, a human practitioner can be tired, stressed or temporarily distracted [13] whereas the Machine Intelligence component of the system would not be affected by these factors. On the other hand, a machine system can make spurious classification mistakes, "hallucinations" [14] that are easily detected by a human specialist.

These and other constituent parts of the complementarity of the human and machine intelligences are based on the fact that, while being capable of achieving high degree of accuracy in recognition tasks, humans and machine systems learn differently, with different data, in quite different ways and processes, and so on, as described in the **Figure 2** below.

Intelligence	Training	Operation
Human	▪ Focus on generalization ▪ Smaller number of cases: patterns and associations, generalization ▪ Longer time to competence (learning curve) ▪ Consultation and collaboration	▪ Influenced by environment and personal factors ▪ Influenced by learning quality, competence ▪ Training through experience ▪ Incremental approach to solution ▪ Association and creativity
Machine Intelligence (conventional models)	• Needs massive sets of annotated cases • Depends on high confidence annotations (trusted prior knowledge) • Learning via informative (differentiating) features • Shorter training time	• Stable performance with respect to environment and individual factors • Cannot easily correct wrong decision • Explainability challenge • Limited creativity

Figure 2. Essential differences between human and machine intelligences.

Based on the observations discussed above, obtaining the decision inputs of both human practitioner and a machine intelligence system (Automated Recognition System) and combining them in producing the diagnostic decision can help in detecting possible errors of the either component, and improve, to a significant extent according to the published research [11], the outcome of the collaborative decision.

To this end, in the proposed approach, a collaborative AI-human intelligent system with parallel decision channels is envisioned, that is expected to take full advantage of the strengths of the human and machine intelligences while mitigating the chance of error in a combined effort to achieve the best diagnostic outcome. This is achieved by a multi-channel concurrent intelligent decision process that mitigates the probability of a decision error to the second power of the probability of error in a conventional diagnostic workflow.

This solution is illustrated in a possible architecture of a collaborative decision system with parallel human and machine channels that make independent decisions on the provided inputs, for an example in our case, diagnostic images. Considering the diagram of the high-level architecture of such a system illustrated in **Figure 3**, one can observe that the following scenarios are possible in the first, concurrent phase of the collaborative system:

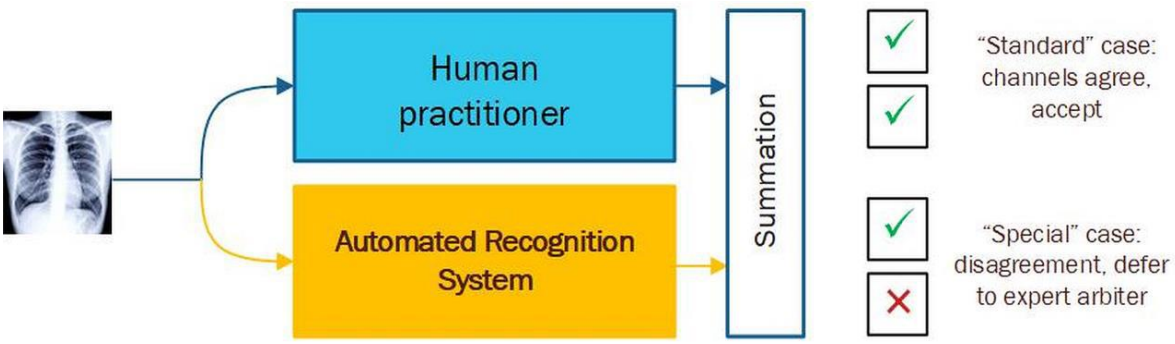


Figure 3. Collaborative AI-human intelligent decision system with concurrent channels.

A. Both human specialist and the machine system produce an erroneous decision. It could be a complex or rare case, where could struggle and err simultaneously or a random coincidence of errors of the components (channels). In that case, one can expect that due to complementary nature of human and machine intelligences discussed above, this possibility would be suppressed as roughly the multiple of the failure (i.e., error) rates of the components. This case would produce a wrong diagnostic decision, and a diagnostic error.

B. The channels agree on a correct decision. The decision is accepted in as the final one with no need for second opinion or verification.

C. This scenario can transpire if the initial decisions of the channels "disagree"; then the final decision is referred to a human expert in the diagnostic area. This case can produce an erroneous result only if the expert and one of the channels err simultaneously (i.e., on the same case, input); then, the error of the final decision will be strongly suppressed in this case as well.

As can be concluded, in this case an error of the final diagnostic decision will be suppressed at least to the second power of the characteristic error of an individual channel. The benefit of the collaborative process comes afore most in the case of mixed decisions, whereby erroneous decisions by individual channels are "caught" by the concurrent channel and referred to an expert arbiter in the area/condition (**Figure 4**).

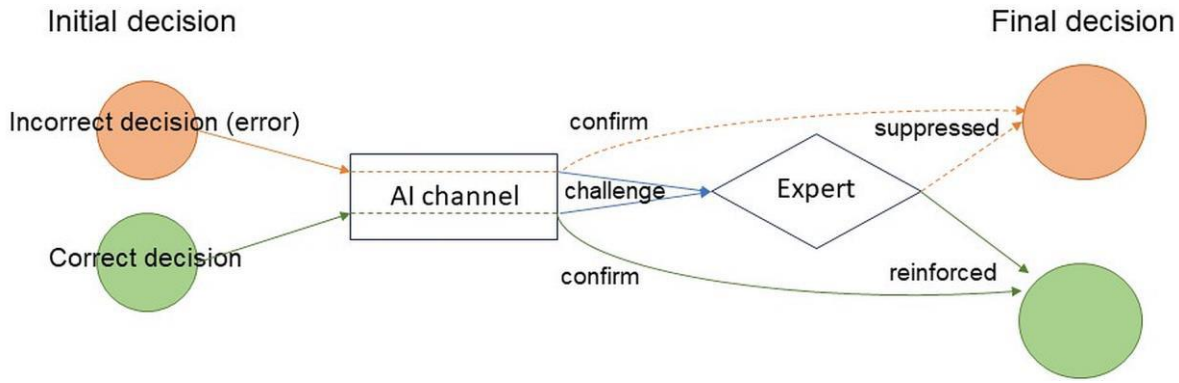


Figure 4. Reducing diagnostic error via intelligent collaboration of concurrent decision channels.

The analysis above demonstrates that integration of a concurrent machine intelligence system can indeed improve the outcome of essential and critical decision via a mechanism of suppression of routine errors and reinforcement of correct decisions.

In contrast, in the conventional "single link" diagnostic workflow, the possibility of an erroneous diagnostic decision is not suppressed by any mechanism and it may not be detected until much later in the treatment cycle [15].

The proposed system would have a potential to eliminate many or most of routine errors that can account for majority of diagnostic errors and can have the following advantages, compared to conventional "single chain" diagnostics workflows.

- Quality: significant improvement in overall accuracy of diagnostics decisions and associated reduction in the follow up spending in the system, improved patient care and overall quality.

- Performance: improves throughput of the diagnostic tasks through the system; balanced and scalable operational model, fully compatible with distributed, high performance and outstanding quality operational models of public service delivery.
- Stimulate optimal use of the expert resources only in those cases that require their attention and expertise.
- Safety and trust: retained full human control over the diagnostic decision.
- Cost efficiency: small cost of development, deployment and operation compared to massive cost saving in the health system to due reduced error and improved outcomes.
- Flexibility: the solution is adaptable to different conditions and types of input data and can be transferred to different areas of integration of Machine Intelligence in medical applications, as well as collaborative decision systems in other areas of application.

The potential for improvement described above can be attributed to the emerging opportunity to combine the strengths and advantages of human and machine intelligences for a significant improvement in the quality of diagnostic decisions over the current practice, while retaining complete and uncompromised human control over the process of diagnostics and treatment.

Moreover, another important advantage of the proposed collaborative system is the potential for continuous improvement. Indeed, cases that resulted in eventual diagnostic errors of all discussed types can be reviewed by the experts in the diagnostic area/condition and integrated into the training processes for human specialists and machine systems. For example, expanding training sets with new samples, drawing attention to certain cases in instruction of practitioners) to attain further gain of quality in each new iteration of the system.

Thus, not only the proposed system can be expected to achieve a substantial improvement in quality at the time of release; but it can be integrated naturally into a lasting process of continuous iterative quality improvement over the lifecycle of the solution.

A combination of high-performance diagnostic AI models with intelligent multi-channel decision system incorporating human in the loop for maximum safety in the operational practice can produce, according to the published results [11] a significant improvement in the accuracy of image diagnostic, and correspondingly significantly reduce the negative impacts and cost of misdiagnosis in the evaluated conditions and areas of public healthcare.

4. Further opportunities for integration of AI in image diagnostic

A practical approach to planning of large-scale integration of collaborative AI-assisted decision systems and models in the practice of diagnostics can include a blueprint proposal for a regional or multi-national anonymous database of diagnostic images that can be used in the research, including development of descriptive core image generative and domain models that can be adapted to specific diagnostic conditions and requirements with minimal effort and lead time. One of many such opportunities can be based on the well-researched practice of transfer learning in the

area of image recognition. Based on availability of dataset(s) of representative samples and the verified methods in transfer learning, effective diagnostic systems for a broad range of conditions can be developed in in a short time, based on established and verified general framework.

Another promising approach being widely developed these days is the development of specific large domain models in diagnostic imaging. Due to limitations of the format, these research directions and approaches will be discussed in more detail elsewhere.

5. Practical implementation of collaborative decision systems

Projects in practical development of collaborative intelligent decision systems can comprise the following key phases and activities.

- A comprehensive review of the current state of the art in image processing, classification and recognition including in the domain of medical image diagnostics and specific medical conditions.
- Research, review, collection, compilation and acquisition of sufficient, by size, representativity, etc., datasets of diagnostic images.
- Research and development of image recognition models, generative and large domain models, preprocessing and other methods for the AI-based component of collaborative models.
- Prototype implementation of the collaborative decision system with realistic operational characteristics.
- Verification, corrections, adjustment and optimization of the collaborative decision system.
- Reviews, information exchanges, presentations, demonstrations and discussions with the practitioner and expert specialists in the field.
- Test deployments, collection of feedback, further improvements, tuning and optimizations for mass-scale integration.
- Integration of the tracking information processes and continuous improvement processes, procedures and policies.

6. Conclusions

To tackle the actual and increasingly pressing challenge of the volume, cost and quality of diagnostic decisions in modern public health systems, integration of Machine Intelligence appears to be an obvious direction to the solution. As we discussed in this work, the perception of simplicity in this program can be misguided and lead to unexpected and unwanted consequences.

To avoid them, we first formulated essential expectations and requirements for intelligent systems with integrated AI components, in both technical and social domains, including, importantly, critical matters of explainability and trust.

An approach to development of collaborative human-AI intelligent decision systems discussed here proposes a framework for collaborative decision process that taps into the strong side of each type of intelligence while mitigating their respective downsides and weaknesses. As a result, a noticeable improvement in the outcome, measured by the overall diagnostic error can be expected in the domains and

conditions where effective integration of machine intelligence is possible and warranted. An additional positive effect of the discussed framework and architecture is the complete traceability of the process that allows effective and straightforward initiation of continuous improvement feedback loops.

The authors expect that the findings and the discussion presented in this work will be of benefit to the research and general community and will contribute to the program of development of performant, accurate, transparent and responsible collaborative intelligent systems for the ultimate benefit of the society.

Conflict of interest: The author declares no conflict of interest.

References

1. Sheikh A, Donaldson L, Westfall-Bates D, et al. Diagnostics errors. WHO Technical series on safer primary care. Available online: https://apps.who.int/iris/bitstream/handle/10665/252410/9789241511636-eng.pdf (accessed on 20 May 2024).
2. Kondro W. Canadian report quantifies cost of medical errors. Lancet. 2004; 363(9426): 2059.
3. Liu X, Faes L, Kale A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet. 2019; 1(6): 271-297.
4. The Guardian. AI equal with human experts in medical diagnosis, study finds | Artificial intelligence (AI). The Guardian News & Media; 2019.
5. Rodriguez RC, Alaniz S, Akata Z. Modeling conceptual understanding in image reference games. In: Advances in Neural Information Processing Systems, Vancouver, Canada. 2019. pp. 13155–13165.
6. Dolgikh S. Low-dimensional representations in unsupervised generative models. In: 20th International Conference Information Technologies-Applications and Theory (ITAT 2020), Slovakia CEUR-WS.org 2718. 2021. pp. 239–245.
7. Bartneck C, Lütge C, Wagner A, Welsh S. Trust and fairness in AI systems. In: An Introduction to Ethics in Robotics and AI, SpringerBriefs in Ethics. Springer, Cham; 2020.
8. Bogen M. All the ways hiring algorithms can introduce bias. Harvard Business Review. 2019.
9. Wang D, Churchill E, Maes P, et al. From Human-Human collaboration to Human-AI collaboration: designing AI systems that can work together with people. In: CHI Conference on Human Factors in Computing Systems (CHI ‘20). 2020. pp. 1-6.
10. Papadopoulos GTh, Antona M, Stephanidis C. Towards Open and Expandable Cognitive AI Architectures for Large-Scale Multi-Agent Human-Robot Collaborative Learning. IEEE Access. 2021; 9: 73890-73909. doi: 10.1109/access.2021.3080517
11. Dolgikh S. A Collaborative Model for Integration of Artificial Intelligence in Primary Care. Journal of Human, Earth, and Future. 2021; 2(4): 395-403. doi: 10.28991/hef-2021-02-04-07
12. Kulikowski CA. Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art—with Reflections on Present AIM Challenges. Yearbook of Medical Informatics. 2019; 28(01): 249-256. doi: 10.1055/s-0039-1677895
13. Sidorov P. The burnout syndrome in communicative professional workers. Gigiena i San-itariia. 2008; 3: 29-33.
14. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 2023; 55(12): 1-38. doi: 10.1145/3571730
15. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Family Practice. 2008; 25(6): 400-413. doi: 10.1093/fampra/cmn071

Imaging and Radiation Research 2024, 7(1), 7128.
https://doi.org/10.24294/irr7128

Article

COVID-19 lesions image segmentation method based on UniFormer

Peng Geng[1,*], Ziye Tan[1], Xiao Cao[2], Xiao Wang[1], Yimeng Wang[1], Dongxin Zhao[1], Conghe Wang[1]

[1] School of Information and Science Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[2] Hebei Hua Zheng Information Engineering Co., Ltd., Shijiazhuang 050043, China
*** Corresponding author:** Peng Geng, Gengpeng@stdu.edu.cn

CITATION

Geng P, Tan Z, Cao X, et al. COVID-19 lesions image segmentation method based on UniFormer. Imaging and Radiation Research. 2024; 7(1): 7128. https://doi.org/10.24294/irr7128

ARTICLE INFO

Received: 13 April 2024
Accepted: 29 May 2024
Available online: 10 June 2024

Abstract: In view of the fact that the convolution neural network segmentation method lacks to capture the global dependency of infected areas in COVID-19 images, which is not conducive to the complete segmentation of scattered lesion areas, this paper proposes a COVID-19 lesion segmentation method UniUNet based on UniFormer with its strong ability to capture global dependency. Firstly, a U-shaped encoder-decoder structure based on UniFormer is designed, which can enhance the cooperation ability of local and global relations. Secondly, Swin spatial pyramid pooling module is introduced to compensate the influence of spatial resolution reduction in the encoder process and generate multi-scale representation. Multi-scale attention gate is introduced at the skip connection to suppress redundant features and enhance important features. Experiment results show that, compared with the other four methods, the proposed model achieves better results in Dice, loU and Recall on COVID-19-CT-Seg and CC-CCIII dataset, and achieves a more complete segmentation of the lesion area.

Keywords: convolutional neural network; COVID-19 lesion image segmentation; self-attention mechanism; multiscale attention gate; spatial pyramid pooling

1. Introduction

Since the end of 2019, the COVID-19 pandemic has affected all aspects of human life. COVID-19 causes multiple issues, including dry cough, fever, headache, myalgia and chest troubles [1]. The diagnostic methods for COVID-19 using medical imaging technology mainly include computed tomography (CT) [2], magnetic resonance imaging (MRI) [3], and X-ray [4]. Compared with X-ray scanning, CT images has higher resolution and higher contrast and are better than X-ray images in displaying soft tissues and small lesions [5]. COVID-19 image segmentation can be introduced to accurately diagnose diseases and provide important information for doctors [6]. However, COVID-19 image segmentation requires experienced radiologists to complete. When faced with a large number of COVID-19 CT images, the manual lesions segmentation consumes a lot of time and is labor-intensive. At the same time, the results of lesions segmentation are easily affected by the radiologist's experience. These subjective and objective factors may lead to large deviations in COVID-19 CT images segmentation [7–9]. Therefore, it is necessary to design robust and accurate COVID-19 CT image segmentation method.

In COVID-19 lung image segmentation, UNet is a commonly used lung region and lesion segmentation. Milletari et al. [9] proposed V-Net in which residual blocks are used as basic convolutional blocks. On this basis, due to the infected area with low contrast in COVID-19 images and the large differences in the infected areas of different patients, accurate segmentation of the infected area is very challenging. In response to the slight differences among healthy tissues. infected tissues and noise,

Wang et al. [10] proposed COPLE-NET for segmenting COVID-19 infected lesions, which introduced maximum pooling and average pooling together in the encoding stage.

Compared with UNet, UNet++ [11] has the advantage of capturing different levels of features. Wang et al. [12] proposed a two-stage method for separating lesions from the lung COVID-19 CT images based on UNet++. In order to segment out more complete structures and more accurate detail information, Cong et al. [13] proposed an end-to-end COVID-19 infection segmentation network

Although the importance of boundary feauture is taken into account in BSNet, it is difficult to segment infected areas because the lesions on the chest CT images are scattered and it is difficult to obtain global semantic information for CNN-based methods. Ibtehaz et al. [14] modified skip connections and the convolution blocks in the ordinary UNet to strengthen the ability of long-range dependencies and multi-level feature combination.

In summary, although CNN-based method above has extraordinary feature representation capabilities, its limited receptive field limiting the accuracy of COVID-19 CT image segmentation. Zhou et al. [15] used the encoder of U-Net to obtain feature representation, input the feature representation of each layer into the attention mechanism, reweighted along the channel direction and spatial direction to obtain the most informative representation, and finally obtain the segmentation result through the decoder. Li et al. [16] aimed at the shortcomings of COVID-19 image segmentation method, such as low contrast between ground-glass opacity and background, blurred boundaries, and difficulty in accurate segmentation. A reverse attention module [13] was added to the skip connection of Unet as a fine marker to identify infected areas in the cleaning strategy. This method can learn the details of the complementary areas and focus on the segmentation of the boundary areas. Xiao et al. [17] proposed a new improved UNet++ model, in which squeeze and excitation attention blocks are adopted to adjust channel of the feature map. The the weights of task-related pixels are strengthened and the background and noise are suppressed. Zhao et al. proposed a Unet++ variant architecture SCOATNet [18] where a new spatial and channel attention module are proposed. The attention mechanism helps to enhance the weight of the infected area in the COVID-19 image and suppress interference in non-lesion areas, thereby raising the accuracy of COVID-19 image segmentation. However, the attention mechanism model fails to capture global dependency because of its smaller receptive field [19]. Therefore, the lesion area in the COVID-19 image cannot be perceived in the global scope, making the effect of the COVID-19 image segmentation poor. To overcome these difficulties, this paper designs a new image segmentation method UniUNet based on UniFormer which has better capabilities of capturing global information. The main contributions of this paper are list as following:

(1) A U-shaped network structure based on UniFormer is proposed, which can effectively remove local redundant information of adjacent slices, and build dependencies on distant lesion areas to improve the accuracy of lesion areas segmentation.

(2) Aimed to capture information of different scales, the Swin spatial pyramid pooling module is introduced into the bottom of the encoder. The module captures the global and local features of lesions through windows with different scales, forming a

multi-scale feature representation, which is helpful to strengthen the ability to capture COVID-19 lesions features.

(3) Multi-scale attention gate is introduced into each skip connection. The module selects features adaptively through convolution of three different receptive fields, and selects valuable features through point convolution voting to further improve the segmentation effect.

2. UniUNet network structure

The proposed UniUNet in this paper is shown in **Figure 1**, which uses the UniFormer blocks instead of the traditional convolution blocks to form a symmetric encoder-decoder structure with skip connection. Firstly, the COVID-19 lesion images are divided into image blocks which are input into the encoder. In the encoder, UniFormer blocks are introduced to carry out local to global self-attention processing on the feature maps. Swin spatial pyramid pooling module is added at the bottom of the encoder to capture features of different scales. The decoder uses the extension layer of the image block for up-sampling, and fuses the multi-scale features from the encoder through the skip connection structure with multi-scale attention gate to enhance the valuable features. Finally, the width and height of the feature maps is restored by linear mapping, and the pixel-level COVID-19 image segmentation is realized.

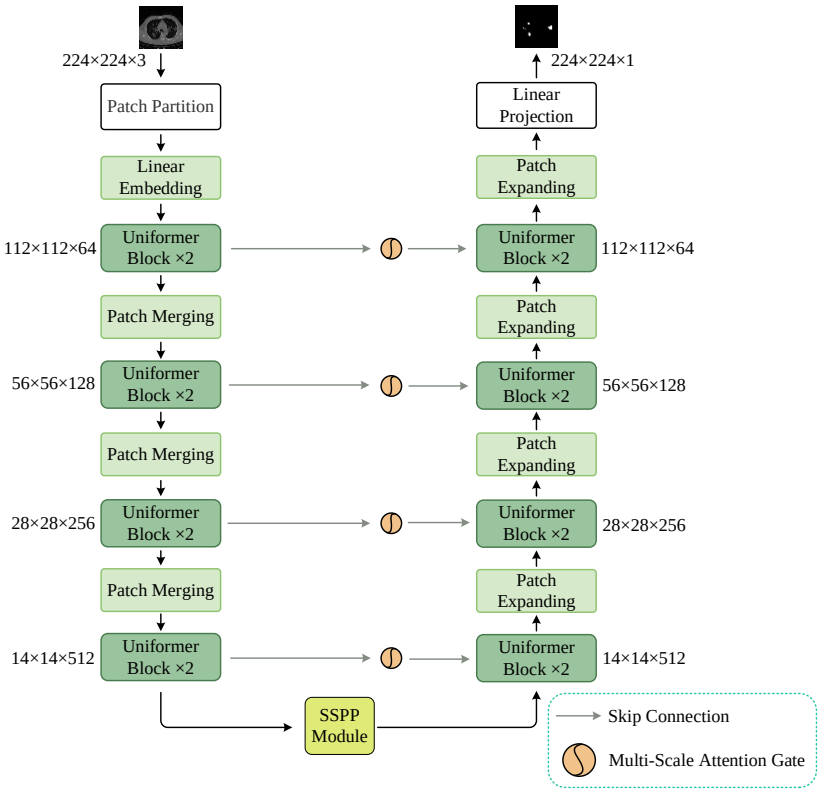


Figure 1. UniUNet model structure.

Patch Merging layer is used reduce resolution of feature maps with down sampling, adjusting the number of channels, and saving computation while maintaining information integrity. Unlike traditional pooling operations, Patch

Merging achieves down sampling by concatenating adjacent patches, ultimately reducing the number of channels through linear layers. In contrast to patch merging, a Patch Expanding layer is adopted to enlarge the resolution of feature maps to twice b up-sampling operation. The implementation process of Patch Merging and Patch Expanding can refer to reference [20].

2.1. UniFormer

To address the issue of local redundancy and global dependence in video, Li et al. [21] proposed UniFormer, which seamlessly integrated the advantages of spatio-temporal self-attention and three-dimensional convolution, A good balance between computational complexity and accuracy has been achieved. The aggregator learns local relations in the shallow layer through a small learnable parameter matrix, and learns global relations in the deep layer through similarity comparison, which effectively balances the computation cost and accuracy. Although UniFormer was originally designed for video processing, this study applied it to two-dimension COVID-19 CT image segmentation to solve the local similarity and the global dependence between infected areas, thus improving the segmentation accuracy. Specifically, the UniFormer module is composed of; three key parts: Multi-Head Relation Aggregator (*MHRA*), Dynamic Position Embedding (*DPE*) and Feed-Forward Network (*FFN*). The UniFormer module formula is as follows:

$$X = DPE(X_{in}) + X_{in} \tag{1}$$

$$Y = MHRA(Norm(X)) + X \tag{2}$$

$$Z = FFN(Norm(Y)) + Y \tag{3}$$

2.2. Swin spatial pyramid pool module

The encoder of UniUNet model includes UniFormer block and patch merging layer, but this will lead to the decrease of spatial resolution. Research [22] suggests that multiscale strucute is particularly effective for extracting high-level feature. In order to strengthen the spatial representation and the multi-scale representation, the Swin spatial pyramid pooling [23] module (SSPP) is inserted at the bottom of the encoder, as shown in **Figure 2**, which captures the global and local features of COVID-19 images with different windows to form the multi-scale representation. Then, these features are sent to the Cross Contextual Attention module for nonlinear fusion to capture a more comprehensive image representation.

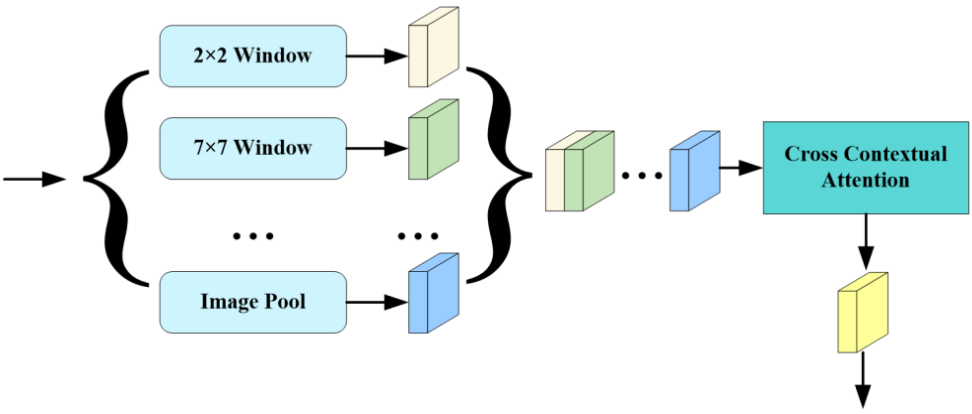


Figure 2. Swin spatial pyramid pooling module.

2.3. Multi-scale attention gate

In the original UNet network structure, not every layer of encoder features can output useful features after skip connection to help the network segment the lesion area. For different datasets and different parameter settings, some redundant information and noise may be generated, which may lead to incorrect segmentation on the infected area of the COVID-19 image, thus affecting the segmentation results of the network model. In order to suppress unimportant and redundant information and enhance valuable features during the skip connection process, this paper introduces the Multi-Scale Attention Gate (MSAG) module [24] in the skip connection part of the structure, as shown in **Figure 3**. In order to adaptively select features with different resolutions, Pointwise convolution, Standard convolution, and dilated convolution combine are combined to extract features with different receptive fields. Each convolution has a batch normalization layer. The feature maps generated by the three kinds of convolutions are of the same size. Before ReLU activation function, the feature maps are connected. Another point convolution is used to capture the important features.

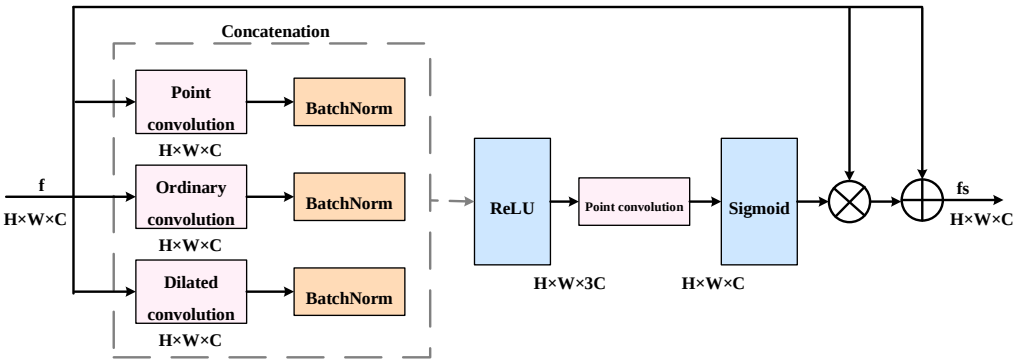


Figure 3. Multi-scale attention gate module.

3. Experimental setting and evaluation metrics

3.1. Dataset

To verify the effect of the proposed UniUNet, we used two publicly\COVID-19

datasets, namely the CC-CCII dataset [25] and the COVID-19-CT-Seg [26] dataset. The COVID-19-CT-Seg dataset is made up of 20 labeled COVID-19 CT sequences captured from 20 patients. The leision areas were labeled by two radiologists and verified by radiologists. The CT resolution is 512 × 512 and 680 × 680. There are 3320 slices in this dataset and 1843 slices exits lesion. We randomly selected 1394 slices from 14 patients for training and testing. The 70% slices are used as training sets and the other is set as test sets. All images were resized to 224 × 224, and all slices of each patient are only used as training or test. In CC-CCII segmentation dataset, there are 750 slices from 150 COVID-19 patients. The background, lung area, ground-glass opacity, and consolidation are manually labled in all of slices. It is worth noting that only consolidation areas and ground-glass opacity are included. There are 540 images with infected areas in CC-CCII dataset. All of images in CC-CCII dataset are randomly separated into training sets and test sets according to the ratio 7:3.

3.2. Experiment settings

All of the resolution is adjusted to 224 × 224. NVIDIA GeForce GTX 1660 SUPER GPU is used for training. The RMSProp optimizer is used. The weight decay and momentum are set to 1e-8 and 0.9, respectively. The two different datasets are trained for 100 and 300 epochs, respectively, the batchsize and initial learning rate were set to 6 and 0.000125, respectively. In addition, the early stopping strategy is used to train the proposed UniUnet. The training will be terminated if the metrics did not increase within 30 epoches. During the process of training, the binary cross entropy loss (BCE Loss) has the advantages of low calculation cost, and fast convergence speed, and is good at the task of binary classification. However, the BCE Loss is limited in effectiveness in dealing with class imbalance issue. The Dice Loss function has strong ability to unbalanced task, but it may be instable to segment small lesion areas in the image. In most COVID-19 images, the lesion are usually scattered and small, which is the issue of small object segmentation. Hybird Loss of Dice Loss [27] and BCE Loss [28] can effectively avoid these disadvantages and focus on detail information more stably. Therefore, the Dice loss function work together with the binary cross entropy to train the proposed UniUnet model. The hybrid loss is expressed as:

$$Loss = \alpha L_{bce} + L_{dice} \tag{4}$$

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}[(1 - y_i)\, log(\,1 - p_i) + y_i\, log\, p_i] \tag{5}$$

$$L_{dice} = 1 - \frac{\sum_{i=1}^{t} y_i p_i + \varepsilon}{\sum_{i=1}^{t} y_i + p_i + \varepsilon} \tag{6}$$

3.3. Evaluation metrics

Six metrics are used to quantitatively, evaluate the effect of different methods. They are Dice, IoU, Accuracy, Precision, Recall, and Specificity [29,30]. The definitions of the above important indicators are as follows:

(1) IoU is Intersection-over-Union ratio: also known as the Jaccard index which is one of the most commonly used indicators in medical image segmentation. It is the ratio of the overlapping area between the Ground truth and the segmentation result to the union area between them.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{7}$$

Dice is same expression formula as $F1 - Score$ is used to evaluate the similarity between the Ground truth and the segmentation result. The larger the value of Dice, the closer the algorithm segmentation result is to the Ground truth, and the better the segmentation effect.

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

(2) Recall is used to indicate the ratio of the number of pixels in the lesion area correctly classified by the network structure to the total number of pixels in the lesion area.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

(3) Precision is used to indicate the ratio of the pixels of the lesion correctly classified by the network model to the total pixels of the lesion in the prediction result.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

(4) Accuracy is used to indicate the ratio of the pixels in the infected area and the pixels in the non-infected area accurately classified by the model to the all of pixels in the image. The higher the accuracy, the better the performance of the segmentation algorithm.

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \tag{11}$$

Specificity indicates the ratio of the pixels in non-infected area correctly classified by the model to all of pixels in non-lesion area. The higher the Specificity is the better the network can distinguish between lesion areas and non-lesion areas.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

where *TP* represents True Positive which represents that the actual sample is positive and the predicted result is positive. *FP* represents False Positive which represents that the actual sample is negative, but the predicted result is positive. *TN* represents True Negative which means that actual sample is negative ads the predicted result is negative. *FN* represents False Negative which means that the actual sample is positive, but the predicted result is negative.

4. Experiment results and analysis

4.1. Ablation experiment

To verify the effectiveness of the SSPP module in the model in capturing multi-scale representation and the MSAG module in suppressing unimportant information, this section completes the ablation experiment of each module on the CC-CCII dataset. As can be seen from **Table 1**, this section uses the network structure without the SSPP module and the MASG module as the backbone network. With the SSPP module, the Dice index is improved by 3.72%, indicating that adding multi-scale representation to the network is effective, and multi-scale context extraction is conducive to capturing the global information of each lesion block in the image and the details on each lesion edge; after adding the MASG module, except for some decreases in accuracy, precision and specificity, other indicators have increased to varying degrees, among which the Dice value, IoU value and Recall value increased by 1.44%, 1.78% and 6.31%, respectively, which further proves that adding MSAG to the skip connection can suppress noise and enhance valuable features, so that the network retains more valuable global and small lesion areas during decoding. The ablation experiment results show that both SSPP and MSAG blocks can effectively increase the COVID-19 CT image segmentation performance.

In addition, FLOPsand parameter after embedding SSPP and MASG. 1 MASG, 2 MSAG, 3 MSAG, and 4 MSAG represent the with different number of MSAG moudles from bottom to top in the skip connection shown in **Figure 1**. For example, Backbone+SSPP+1 MSAG represents one MSAG module is embedded in the bottom skip connection in **Figure 2**. Backbone + SSPP + 2 MSAG represents that the other MSAG module is inserted tin the second skip connection from bottom to top in **Figure 2** based on Backbone + SSPP + 1 MSAG. From **Table 2**, it can be seen that the SSPP module has significantly increased FLops and Parmeters. Only one MSAG module resulted in a slight increase in FLops. However, adding four MSAG modules significantly increases the parameter count and FLOPS compared to only SSPP modules.

Table 1. Ablation experiment on CC-CCII dataset (%).

Methods	**Dice**	**Recall**	**IoU**	**Precision**	**Specificity**	**Accuracy**
Backbone	76.22	73.52	62.12	79.96	99.68	99.25
Backbone + SSPP	79.94	76.71	66.88	83.98	99.75	99.37
Backbone + SSPP + MSAG (UniUNet)	81.38	83.02	68.75	80.17	99.63	99.36

Table 2. FLOPs and Params with MSAG or not.

Methods	FLOPs(G)	Params(M)
Backbone	55.74	9.75
Backbone + SSPP	82.92	33.56
Backbone + SSPP + 1 MSAG	89.78	33.65
Backbone + SSPP + 2 MSAG	96.60	34.01
Backbone + SSPP + 3 MSAG	103.40	35.46
Backbone + SSPP + 4 MSAG	110.19	41.23

4.2. Comparison with other methods

Aim to prove the effectiveness of the proposed UniUNet, two COVID-19 image

segmentation methods and two medical image segmentation methods are compared. Infnet [31] is a COVID-19 image segmentation method and has been widely used as the baseline for COVID-19 image segmentation. BSNet [13] is based on modeling semantic relationship and guidance of boundary detail to segment lesion area more completely. TFCNs [32] is a state-of-the-art medical image segmentation method constructed using ResLinear Transformer and convolutional neural networks. Swin-Unet [20] is an effective medical image segmentation method using transformer technology to construct a U-shaped structure similar to the proposed UniUNet. Swin-Unet was widely thought as the baseline for medical image segmentation.

(1) Comparison on COVID-19-CT-seg dataset.

As can be seen from **Table 3**, on COVID-19-CT-Seg dataset, UniUNet is superior to Swin-Unet in all evaluation indexes, indicating that it has better training effect on small datasets and can capture the global dependence of lesion areas more effectively. Swin Unet has the lowest Dice, IoU and Precision among these five methods. The Dice, IOU, Recall and Precision of InfNet and BSNet method are relatively closer and better than Swin-Unet. But they are all lower than the proposed UniUNet in this work. Compared with the classic InfNet, UniUNet has improved Dice, Recall and IoU by 3.68%, 5.72% and 4.39% respectively, achieving better results.

In order to analyze the segmentation results of each method in a visual way, **Figure 4** demonstrates the comparison on predicted lesion area with different model. **Figure 4** illustrates that the InfNet cannot efficaciously obtain the edge details of the lesions which is an important basis for doctors to diagnose COVID-19. For the second column and the third column, UniUNet was able to accurately locate several lesion areas in the images, indicating that the proposed UniUNet can better establish the global relationship in the lesion areas. For the upper left lesion area in the fifth column, InfNet and BSNet did not segment the whole lesion area, while TFCNs segmented only part of lesion and showed obvious over segmentation. Compared with other methods, UniUNet is closest to Ground truth in segmentation accuracy, which proves the advantages of the method based on Transformer in COVID-19 lesion segmentation.

Table 3. Quantitative comparison on COVID-19-CT-Seg dataset (%).

Methods	Dice	IoU	Recall	Precision	FLOPs(G)	Params(M)
InfNet	74.01	60.22	78.83	74.14	31.52	30.91
TFCNs	73.22	58.91	74.73	76.02	168.40	105.79
BSNet	74.97	61.94	79.11	76.20	210.39	43.99
Swin-Unet	69.39	54.63	76.11	67.26	35.46	27.14
UniUNet	77.69	64.61	84.55	74.37	110.19	41.23

In addition, **Table 3** also provides a comparison of FLOPs and parameters for different methods. In comparison with the TFCNs, the FLOPs and parameters of the UniUNet method are remarkablely decreased. Compared with BSNet, the FLOPs of the UniUNet method are significantly reduced. The FLOPs and parameters of the UniUNet method are significantly higher than those of InfNet and Swin Unet. However, compared to these four methods, it can be seen from **Figure 4** and **Table 3** that UniUNet significantly improves segmentation accuracy.

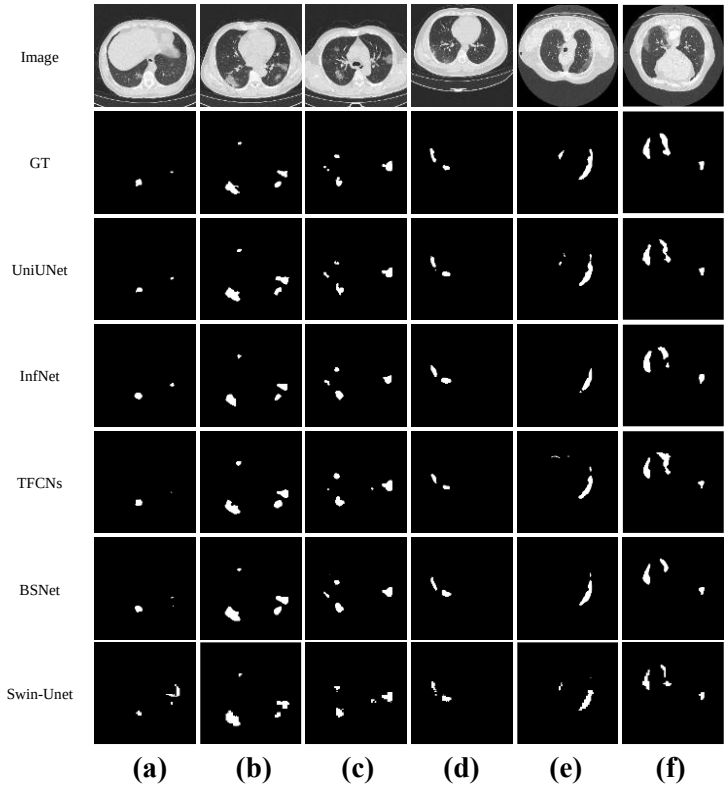


Figure 4. Comparison of segmentation results in COVID-19-CT-Seg dataset.

(2) Comparison on CC-CCII datasets.

In order to verify the effectiveness of the UniUNet model on CC-CCII dataset, another comparative experiment is carried out on the CC-CCII dataset, and six indicators of different methods are list in **Table 4**. The results in **Table 4** verify that UniUNet outperforms other methods in key indicators such as Dice, IoU and Recall, which are 81.38%, 68.75% and 83.02% respectively. In addition, the second or third best results have been achieved on the indicators of Precision, Accuracy and Specificity. Comprehensive analysis of these indicators shows that UniUNet has excellent segmentation performance in COVID-19 image segmentation task, which proves that it is good at COVID-19 lesion segmentation.

Table 4. Metrics Comparison on CC-CCII dataset (%).

Methods	Dice	IoU	Recall	Precision	Accuracy	Specificity
InfNet	77.28	63.12	75.03	80.55	99.27	99.68
TFCNs	76.09	61.73	78.09	74.64	99.18	99.53
BSNet	81.07	68.33	79.73	83.10	99.39	99.72
Swin-Unet	65.68	49.34	70.15	62.77	98.81	99.29
UniUNet	81.38	68.75	83.02	80.17	99.36	99.63

In addition, a comparison of the segmentation results of each method on the CC-CCII dataset is illustrated in **Figure 5**. **Figure 5** demonstrates that segmented results by the other methods have missing segmentation or over-segmentation. For example, the image in the first row, the InfNet, TFCNs and BSNet methods have obvious

missing segmentation on the lesion area at the lower right, and they have not divided it into connected lesion areas. The Swin-Unet method not only presents zigzag phenomenon in the segmentation result, but also has serious over-segmentation. These phenomenon of missing segmentation or over-segmentation indicates that these methods cannot effectively model the global dependency of the lesion. Only UniUNet can completely and accurately segment the larger lesion area. It shows that the proposed method can more accurately locate each part of the region and effectively remove the noise, which is because the multi-scale attention gate module is added between the skip connection to suppress the noise and enhance the valuable features. However, UniUNet is not sensitive enough to small lesions in the image, resulting in the problem of wrong segmentation.

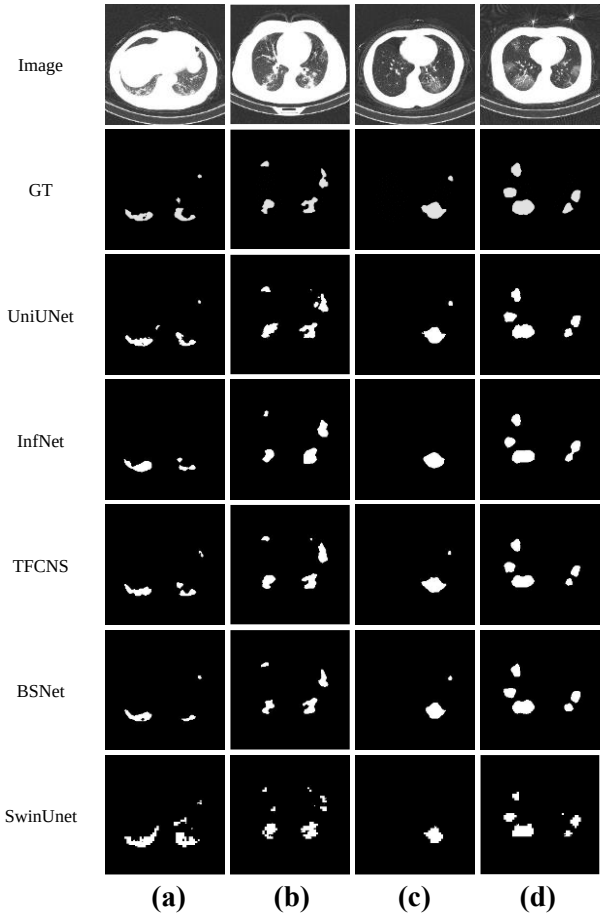


Figure 5. Predicted results with different methods in CC-CCII dataset.

5. Conclusions

Aiming at the problem that the segmentation network based on convolution neural network lacks the global dependence of modeling infected areas in COVID-19 images, which is not conducive to the complete segmentation of scattered lesion areas, a U-shaped COVID-19 image segmentation method based on UniFormer is proposed. UniFormer can establish a good correlation with the global lesion area. SSPP module can obtain multi-scale representation. The MSAG module is used to enhance valuable features in the network. The experiments illustrate that the method UniUnet in this work achieves ideal results on two COVID-19 image datasets, and a more complete

segmentation result is obtained. This method is suitable for training small datasets such as COVID-19 images, and meanwhile, it enhances the dependence of global lesion areas. However, there are still some missing and wrong segmentation phenomena in this method for some small and blurred lesion areas, because Transformer's method pays more attention to global dependence and ignores local details in the image. Therefore, it is particularly important to explore segmentation methods that can effectively retain global and local information.

Author contributions: Conceptualization, PG and ZT; methodology, PG and ZT; software, XC and YW; validation, ZT, PG and DZ; formal analysis, CW; investigation, ZT; resources, PG and ZT; data curation, ZT; writing—original draft preparation, ZT and PG; writing—review and editing, XW; visualization, ZT; supervision, PG; project administration, PG and XC; funding acquisition, PG. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Geng P, Tan Z, Wang Y, et al. STCNet: Alternating CNN and improved transformer network for COVID-19 CT image segmentation. Biomedical Signal Processing and Control. 2024; 93: 106205. doi: 10.1016/j.bspc.2024.106205
2. Niitsu H, Mizumoto M, Li Y, et al. Tumor Response on Diagnostic Imaging after Proton Beam Therapy for Hepatocellular Carcinoma. Cancers. 2024; 16(2): 357. doi: 10.3390/cancers16020357
3. Shimizu S, Nakai K, Li Y, et al. Boron Neutron Capture Therapy for Recurrent Glioblastoma Multiforme: Imaging Evaluation of a Case with Long-Term Local Control and Survival. Cureus. 2023. doi: 10.7759/cureus.33898
4. Li S, Mo Y, Li Z. Automated Pneumonia Detection in Chest X-Ray Images Using Deep Learning Model. Innovations in Applied Engineering and Technology. Published online December 12, 2022: 1–6. doi: 10.62836/iaet.vli1.002
5. Bueno C, Barker MD, Orphan VJ. X-Ray Detector Physics and Applications II. Society of Photo Optical; 1993. doi: 10.1117/12.164737
6. Zheng T, Lin F, Li X, et al. Deep learning-enabled fully automated pipeline system for segmentation and classification of single-mass breast lesions using contrast-enhanced mammography: a prospective, multicentre study. eClinicalMedicine. 2023; 58: 101913. doi: 10.1016/j.eclinm.2023.101913
7. Zhang J, Chen D, Ma D, et al. CdcSegNet: Automatic COVID-19 Infection Segmentation from CT Images. IEEE Transactions on Instrumentation and Measurement. 2023; 72: 1–13. doi: 10.1109/tim.2023.3267355
8. Shi F, Wang J, Shi J, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering. 2021; 14: 4–15. doi: 10.1109/rbme.2020.2987975
9. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV). 2016. doi: 10.1109/3dv.2016.79
10. Wang G, Liu X, Li C, et al. A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images. IEEE Transactions on Medical Imaging. 2020; 39(8): 2653–2663. doi: 10.1109/tmi.2020.3000314
11. Yu L, Hu Z, Zhang F, et al. Unmanned aerial vehicle image biological soil crust recognition based on UNet++. International Journal of Remote Sensing. 2022; 43(7): 2660–2676. doi: 10.1080/01431161.2022.2066486
12. Wang B, Jin S, Yan Q, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. Applied Soft Computing. 2021; 98: 106897. doi: 10.1016/j.asoc.2020.106897
13. Cong R, Zhang Y, Yang N, et al. Boundary Guided Semantic Learning for Real-Time COVID-19 Lung Infection Segmentation System. IEEE Transactions on Consumer Electronics. 2022; 68(4): 376–386. doi: 10.1109/tce.2022.3205376
14. Ibtehaz N, Kihara D. Acc-unet: a completely convolutional unet model for the 2020s. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023: 692-702. doi: 10.48550/arXiv.2308.13680

15. Zhou T, Canu S, Ruan S. An automatic COVID-19 CT segmentation network using spatial and channel attention mechanism. ARXIV preprint arXiv:2004.06673, 2020.
16. Li CF, Xu YD, Ding XH, et al. MultiR-Net: A Novel Joint Learning Network for COVID-19 segmentation and classification. Computers in Biology and Medicine. 2022; 144: 105340. doi: 10.1016/j.compbiomed.2022.105340
17. Xiao H, Ran Z, Mabu S, et al. SAUNet++: an automatic segmentation model of COVID-19 lesion from CT slices. The Visual Computer. 2022; 39(6): 2291–2304. doi: 10.1007/s00371-022-02414-4
18. Zhao S, Li Z, Chen Y, et al. SCOAT-Net: A novel network for segmenting COVID-19 lung opacification from CT images. Pattern Recognition. 2021; 119: 108109. doi: 10.1016/j.patcog.2021.108109
19. Jia W, Ma S, Geng P, et al. DT-Net: Joint Dual-Input Transformer and CNN for Retinal Vessel Segmentation. Computers, Materials & Continua. 2023; 76(3): 3393–3411. doi: 10.32604/cmc.2023.040091
20. Karlinsky L, Michaeli T, Nishino K, et al. Computer Vision – ECCV 2022 Workshops. Springer Nature Switzerland; 2023. doi: 10.1007/978-3-031-25066-8
21. Li K, Wang Y, Gao P, et al. Uniformer: unified transformer for efficient spatiotemporal representation learning. ARXIV preprint arXiv:2201.04676. 2022.
22. Bello IM, Zhang K, Su Y, et al. Densely multiscale framework for segmentation of high resolution remote sensing imagery. Computers & Geosciences. 2022; 167: 105196. doi: 10.1016/j.cageo.2022.105196
23. Azad R, Heidari M, Shariatnia M, et al. Transdeeplab: convolution-free transformer-based deeplab v3+ for medical image segmentation. Proceeding of the International Workshop on Predictive Intelligence in Medicine. 2022: 91-102. doi: 10.48550/arXiv.2208.00713
24. Tang F, Wang L, Ning C, et al. CMU-Net: A Strong ConvMixer-based Medical Ultrasound Image Segmentation Network. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). 2023. doi: 10.1109/isbi53787.2023.10230609
25. Zhang K, Liu X, Shen J, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell. 2020; 181(6): 1423–1433.e11. doi: 10.1016/j.cell.2020.04.045
26. Ma J, Wang Y, An X, et al. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. Medical Physics. 2021; 48(3): 1197–1210. doi: 10.1002/mp.14676
27. Liu J, Zhao D, Shen J, et al. HRD-Net: High resolution segmentation network with adaptive learning ability of retinal vessel features. Computers in Biology and Medicine. 2024; 173: 108295. doi: 10.1016/j.compbiomed.2024.108295
28. Geng P, Lu J, Zhang Y, et al. TC-Fuse: A Transformers Fusing CNNs Network for Medical Image Segmentation. Computer Modeling in Engineering & Sciences. 2023; 137(2): 2001–2023. doi: 10.32604/cmes.2023.027127
29. Chen L, Bentley P, Mori K, et al. DRINet for Medical Image Segmentation. IEEE Transactions on Medical Imaging. 2018; 37(11): 2453–2462. doi: 10.1109/tmi.2018.2835303
30. Wang R, Lei T, Cui R, et al. Medical image segmentation using deep learning: A survey. IET Image Processing. 2022; 16(5): 1243–1267. doi: 10.1049/ipr2.12419
31. Fan DP, Zhou T, Ji GP, et al. Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images. IEEE Transactions on Medical Imaging. 2020; 39(8): 2626–2637. doi: 10.1109/tmi.2020.2996645
32. Cao H, Wang Y, Chen J, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision; 2022. doi: 10.48550/arXiv.2105.05537

Imaging and Radiation Research 2024, 7(1), 4546.
https://doi.org/10.24294/irr4546

Article

Differential diagnosis of hepatocellular carcinoma and cirrhotic nodules via radiomics models based on magnetic resonance images

Changdong Ma[1], Changsheng Ma[2,*], Shuang Yu[3,*]

[1] Department of Radiation Therapy, Qilu Hospital of Shandong University, Jinan 250012, China
[2] Department of Radiation Physics, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan 250012, China
[3] Department of Hematology, Qilu Hospital of Shandong University, Jinan 250012, China
*** Corresponding authors:** Changsheng Ma, machangsheng_2000@126.com; Shuang Yu, yushuang@sdu.edu.cn

CITATION

Ma C, Ma C, Yu S. Differential diagnosis of hepatocellular carcinoma and cirrhotic nodules via radiomics models based on magnetic resonance images. Imaging and Radiation Research. 2024; 7(1): 4546. https://doi.org/10.24294/irr4546

ARTICLE INFO

Received: 4 February 2024
Accepted: 9 April 2024
Available online: 30 April 2024

Abstract: Objective: To investigate the value of differential diagnosis of hepatocellular carcinoma (HCC) and cirrhotic nodules via radiomics models based on magnetic resonance images. **Background:** This study is to distinguish hepatocellular carcinoma and cirrhotic nodules using MR-radiomics features extracted from four different phases of MRI images, concluded T1WI, T2WI, T2 SPIR and delay phase of contrast MRI. **Methods:** In this study, the four kind of magnetic resonance images of 23 patients with hepatocellular carcinoma (HCC) were collected. Among them, 12 patients with liver cirrhosis were used to obtain cirrhotic nodules (CN). The dataset was used to extract MR-radiomics features from regions of interest (ROI). The statistical methods of MRradiomics features could distinguish HCC and CN. And the ability of radiomics features between HCC and CN was estimated by receiver operating characteristic curve (ROC). **Results:** A total of 424 radiomics features were extracted from four kind of magnetic resonance images. 86 features in delay phase of contrast MRI，86 features in spir phase of T2WI，86 features in T1WI and 88 features in T2WI showed statistical difference ($p < 0.05$). Among them, the area under the curves (AUC) of these features larger than 0.85 were 58 features in delay phase of contrast MRI, 54 features in spir phase of T2WI, 62 features in T1WI and 57 features in T2WI. **Conclusions:** Radiomics features extracted from MRI images have the potential to distinguish HCC and CN.

Keywords: radiomics features; hepatocellular carcinoma; MRI; cirrhotic

1. Introduction

The differential diagnosis of liver masses is still the current focus. As The primary liver cancer is one of the most common malignant tumors in the clinic, with more than 840,000 new cases per year and above 780,000 death cases per year, which incidence and mortality rate rank seventh and third in all cancers,respectively [1]. In more than 90% of the cases. The subtype of primary liver cancer is hepatocellular carcinoma (HCC) [2],which complicates liver cirrhosis caused by hepatitis C virus (HCV) and hepatitis B virus (HBV) infection [3]. The evolution of HCC is from cirrhotic nodule(CN) to dysplastic nodule(DN) and then to small hepatocellular carcinoma (SHCC), finally to progressed HCC [4]. SHCC also knowen as early hepatocellular carcinoma (eHCC) or subclinical hepatocellular carcinoma,without clearly imaging characterizations and clincial symptoms.The main reason of high mortality rate of HCC is detected so lately that treatment cannot work out effectively [5]. Thus, the sole approach to achieve long-term survival is to detect the tumor at an early stage.

Although biopsy is the gold standard for identifying focal hepatic lesions, it has limitations: a) Biopsy is an invasive examination, which have difficulity in acceptance of patients and repeatability of sample; b) The particularity of liver anatomy makes sampling difficult, appearing false-negative and false-positive results [5,6]; c) When the needle is withdrawn, it have risks to cause bleeding or implant transfer, which affects the subsequent treatment [6]. Fortunately, many researchers have discovered that the imaging features of SHCC and CN have great research value for differential diagnosis. Huang et al. [7] conclude that contrastenhanced ultrasound (CEUS) could be helpful in the differential diagnosis of hepatic malignant and benign lesions ,but dysplastic nodule may manifest with a similar enhancing pattern as that in welldifferentiated small HCC.Also,US images are easily affected by the operator's technical level and gastrointestinal gas. Chen et al. [8] concluded that 64-slice spiral CT can provide more sufficient imaging evidence for the clinical diagnosis of HCC and FNH and effectively identify benign and malignant tumors compared with conventional US examination, which also has high sensitivity in the diagnosis of tiny lesions.Furthmore, Ronot M and other researches [5] have shown that arterial phase hyperenhancement followed by washout on CT or MRI is highly specific.

However, whether CEUS, enhanced CT, it only distinguish CN from SHCC anatomically. With the continuous deepening of research, many researchers have now adviced that MRI functional imaging is useful for distinguishing diagnosis,which has great potential to research.For example, According to a study [9] of hepatocellular carcinoma based on US ,CTand MR images by some people, the sensitivity of MR images in the hepatobiliary stage is the highest. Moreover, the study on the quantitative evaluation of focal hepatic lesions by DWMRI used 4 b values to obtain different ADC images [10]. The results of the study suggest that ADC values can distinguish cavernous hemangioma and liver cysts. The ratio of the ADC value of leision/liver can distinguish HCC and hepatic metastasis, and can provide information to help diagnose focal hepatic lesions with a diameter less than 3 cm. However, these studies still cannot clearly distinguish SHCC from DN.Radiomics is an emerging technology that has developed in recent years. It uses software to extract the texture features of the region of interest(ROI) by delineating it in the image, and performs computer operations to obtain small image parameters that cannot be observed by the human eyes. This research will use the combination of radiomics and magnetic resonance technology to differentiate between DN and HCC.

2. Methods

2.1. Radiomics workflow

The raidomics flow of this study included: (1) images acquisition; (2) feature extraction; (3) data analysis **(Figure 1**).

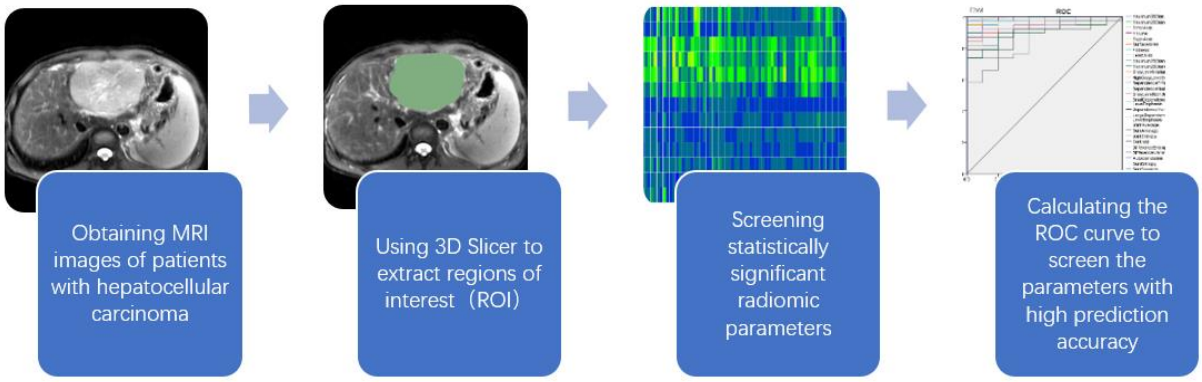


Figure 1. The workflow of the study.

2.2. Patients

The protocol for this study was approved by the Institutional Review Committee of the Shandong First Medical University Affiliated Cancer Hospital Ethics Committee. The ethics filing number is SDTHEC2020010008. Case entry criteria: (1) Complete clinical imaging data; (2) No surgery, radiotherapy, chemotherapy, or interventional treatment before imaging examination; (3) Pathologically confirmed hepatocellular carcinoma.Search for 23 patients with hepatocellular carcinoma in Shandong Cancer Hospital who met the enrollment criteria from April 2019 to January 2020, a total of 24 lesions, and they were recorded as 1 group, including 21 males and 2 females, aged 42–83 years old , An average of 56.08 years old. Among the above-mentioned patients, 12 had a history of liver cirrhosis and hepatitis B, and 12 had cirrhotic nodules, which were recorded as two groups, including 10 males and 2 females.

2.3. Patient images acquisition

Use GE HDe1.5TMR scanner.The scanning sequence and parameters are as follows: Axial breathing trigger FSETWI + FS,TR/TE2-3 breathing cycle/(80 ± 10) ms,layer thickness 6 mm, layer spacing 1.5 mm, field of view (FOV) 40 cm × 36 cm, matrix 320 × 224, number of excitations 2; SE. EPIDWI, TR 5000 ms, TE 75.40 ms, layer thickness 6 mm, layer spacing 1.5 mm, FOV 40 cm × 40 cm, matrix 128 × 128, number of excitations 8; FSPGR TWI inverse phase imaging, TR 120–250 ms, TE 2.25–4.5 ms, layer thickness 6 mm, layer spacing 1.5 mm, FOV 40cm×36 cm, matrix 256 × 170, excitation times 1; Liver Volume Rapid Acquisition (IAVA) three-dimensional dynamic enhancement scan, TR 5.14 ms, TE 2.30 ms,The layer thickness is 5 mm, the layer spacing is 2.50 mm, the FOV is 40 cm × 36 cm, and the matrix is 288 × 192. Using a double-barreled high-pressure syringe, inject Ou Naiying 0.1 mol/kg body weight through the cubital vein at a flow rate of 3 ml/s, and scan the arterial phase, portal vein phase, and equilibrium phase at 18–22 s, 60 s, and 180 s after the contrast agent injection. The size of the liver is about 15–18s to complete a single-phase whole liver scan.

2.4. Region of interests (ROI) segmentation

The images are divided into four categories: T1WI, T2WI, T2 SPIR, and enhanced scan delay period. Two imaging physicians with more than 5 years of work

experience observe all the images separately, and those who have different results discuss and reach an agreement together. Use the imaging omics analysis software 3D slicer 4.8 to delineate the area of interest and obtain 106 children with seven parent features The characteristic data table, the area of interest (ROI)

Of the lesion includes the largest extent of the lesion entity as much as possible, and avoids the blood vessel, hemorrhage, necrosis, and cystic area.Divide the data into four categories:T1WI, T2WI, T2 SPIR, and enhanced scan delay period, and then divide each category into seven groups: Shape, Gldm, Glcm, Firstorder, Glrlm, Glszm, and Ngtdm, and analyze them separately.

3. Statistical analysis

Enter the values of all parameters into html to obtain a heat map representing these data.(**Figure 2**) The statistical analysis software SPSS 22.0 was used to process and analyze the data. Mann-Whitney U test was selected for the imaging omics characteristic parameter data obtained from the MR images of each phase of the cancer and sclerosing nodules to screen

for statistically significant difference parameters between the two lesions. Thus obtained radiological characteristics that can distinguish hepatocellular carcinoma from sclerosing nodules. Then use the ROC curve drawing function in spss to determine the diagnostic performance of the above-mentioned characteristic parameters. The characteristic parameters whose area under the curve is less than 0.85 are eliminated. Thus, imaging characteristics parameters that can efficiently distinguish hepatocellular carcinoma from sclerosing nodules can be obtained.

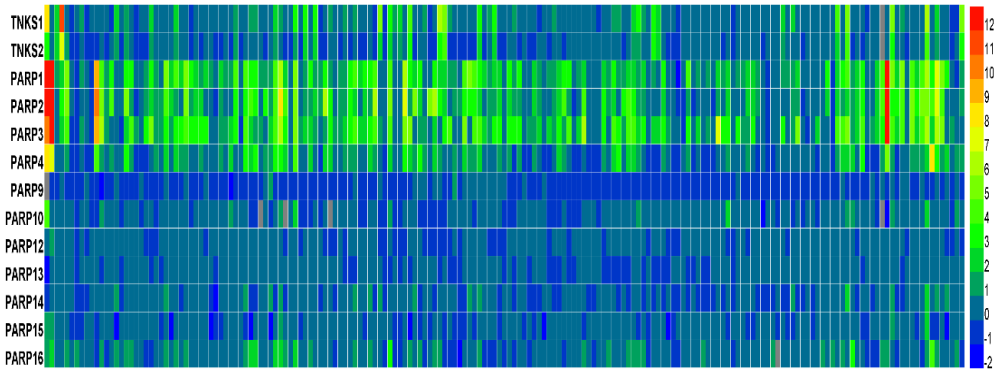


Figure 2. Distribution of all parameters.

3.1. Patient characteristics

In this study, a total of 23 hepatocellularcarcinoma patients were included, including 21 men and 2 women (maximum age 83 years, minimum age 42 years, median age 53 years). Then there are 12 patients with a history of liver cirrhosis among these 23 patients, of which 10 are males and 2 are females (maximum age 66 years, minimum age 42 years, median age 56 years). See **Table 1**.

Table 1. Clinical information of enrolled patients.

Number	Age	Sex	Size (cm)	T1WI	T2WI	T2 SPIR	DELAY	Cirrhosis	Hepatitis
1	62	M	3.0 × 3.0	--	--	√	√	--	--
2	64	M	5.8 × 6.6	--	√	√	√	--	--
3	42	M	2.0 × 1.5	--	√	√	--	positive	HBV
4	63	M	4.6 × 7.0	--	√	√	√	--	HBV
5	49	M	10.2 × 7.6	√	√	√	--	positive	HBV
6	83	M	0.9 × 1.6	√	√	√	--	--	--
7	44	M	4.8 × 6.2	--	√	√	√	positive	HBV
8	56	M	3.3 × 2.5	--	√	√	√	--	HBV
9	58	M	9.3 × 9.0	--	√	√	√	positive	HBV
10	58	M	1.4 × 0.9	--	√	√	--	positive	HBV
11	53	M	2.6 × 2.4	√	√	√	--	positive	HBV
12	62	M	6.9 × 5.0	√	√	√	--	--	HBV
13	49	M	3.2 × 2.9	--	√	√	√	Positive	HBV
14	66	F	2.8 × 2.8	--	√	√	√	Positive	HBV
15	58	F	4.4 × 3.5	√	√	√	--	Positive	HBV
16	48	M	8.0 × 5.1	√	√	√	--	Positive	HBV
17	50	M	2.7 × .3	--	√	√	√	Positive	HBV
18	46	M	14.1 × 9.8	--	√	√	√	-	HBV
19	50	M	8.2 × 8.4	--	√	√	√	Positive	Positive
20	42	M	9.9 × 7.2	√	√	√	--	Positive	positive
21	67	M	7.2 × 6.6	√	√	√	--	-	HBV
22	60	M	10.9 ×8.5	--	√	√	√	Positive	HBV
23	60	M	11.0 × 10.5	√	√	√	--	Positive	HBV

4. Feature results

In this study, a total of 106 imaging radi-omics features of 24 hepatocellular carcinomalesions and 12 sclerosing nod-ules lesions were extracted. According to the imaging omics, these 106 features can be divided into 7 catgories. Be more detailed,shape 13 features, gldm 14 features,glcm 24 fetures, firstorder 18 features,glrlm 15 features,glszm 16 fea-tures, ngtdm 5 features.

4.1. Statistical results

All data have been tested by the Mann-Whitney U test, and the *p*-values obtained are shown in **Table 2**. As shown in **Table 2**, among all four imaging methods, there are 70 types of statistically significant differences in imaging features between hepatocellular carcinoma and cirrhotic nodules. They were 86 features in T1WI and 88 features in T2WI, 86 features in delay phase of con-trast MRI and 86 features in spir phase of T2W.

Table 2. Feature parameters and differentiating between cirrhotic nodules and hepatocellular carcinoma.

Category	Feature	T1wI	P value	DELAY	TIWI	T2WI	ROCSPIR	DELAY
	Maximum3DDiameter	0.001	T2WI	0.044	1	1	SPIR	0.792
	Maximum2DDiameterSlice	0.001	0	0.011	1	1	0.958	0.857
	MinorAxis	0.001	0	0	1	1	0.955	1
	Volume	0.001	0	0	1	1	1	1
shape	MajorAxis	0.001	0	0.104	1	1	1	0.74
	SurfaceArea	0.001	0	0	1	1	0.955	1
	Flatness	0.003	0	0	0.96	0.992	1	1
	LeastAxis	0.001	0	0	1	0.996	1	1
	Maximum2DDiameterColumn	0.001	0	0	1	1	1	1
	Maximum2DDiameterRow	0.001	0	0	1	1	1	1
	GrayLevelVariance	0.003	0	0.011	0.96	0.995	1	0.857
gldm	HighGrayLevelEmphasis	0.001	0	0.004	1	1	0.966	0.896
	DependenceEntropy	0.001	0	0	1	1	0.992	1
	DependenceNonUniformity	0.001	0	0	1	1	1	1
	GrayLevelNonUniformity	0.001	0	0	1	0.963	1	1
	SmallDependenceHighGrayLevelEmphasis	0.594	0.203	0.006	0.6	0.889	0.992	0.883
	LargeDependenceEmphasis	0.008	0	0.002	0.92	0.829	0.777	0.922
	DependenceVariance	0.008	0.001	0	0.92	0.945	0.818	0.974
	LargeDependenceHighGrayLevelEmphasis	0.001	0	0.002	1	1	0.958	0.922
	JointAverage	0.001	0	0.002	1	1	1	0.922
glcm	SumAverage	0.001	0	0.002	1	1	0.996	0.922
	JointEntropy	0.003	0	0	0.96	0.997	0.996	1
	Idmn	0.594	0	0.011	0.4	0.655	0.977	0.857
	Contrast	0.04	0.072	0	0.84	0.934	0.939	0.961
	DifferenceEntropy	0.005	0	0	0.94	0.966	0.886	1
	DifferenceVariance	0.005	0	0	0.94	0.966	0.958	1
	Idn	0.594	0	0.011	0.4	0.582	0.951	0.857
	Correlation	0.594	0.35	0	0.4	0.655	0.773	1
	Autocorrelation	0.001	0.062	0.004	1	1	0.97	0.896
	SumEntropy	0.001	0	0	0.98	1	0.996	1
firstorder	SumSquares	0.003	0	0.006	0.96	0.982	1	0.883
	ClusterProminence	0.001	0	0.008	1	1	0.947	0.87
	Imc2	0.001	0	0.375	0.98	0.645	0.977	0.636
	DifferenceAverage	0.055	0.076	0.002	0.82	0.889	0.61	0.922
	ClusterTendency	0.001	0	0.006	1	0.997	0.833	0.883
	InterquartileRange	0.001	0	0.011	1	0.963	0.966	0.857
	Energy	0.001	0	0	1	1	0.909	1
	RobustMeanAbsoluteDeviation	0.001	0	0.011	1	0.966	0.992	0.857
	MeanAbsoluteDeviation	0.001	0	0.011	1	0.984	0.928	0.857
	TotalEnergy	0.001	0	0	1	1	0.962	1
	Maximum	0.001	0	0	0.88	0.958	0.992	1

Table 2. (*Continued*).

Category	Feature	T1wI	P value	DELAY	TIWI	T2WI	ROCSPIR	DELAY
	90Percentile	0.055	0	0.003	0.44	0.842	0.788	0.909
	Entropy	0.001	0.001	0	0.96	1	0.648	1
	Range	0.003	0	0.002	1	1	0.985	0.922
glszm	Variance	0.001	0	0.011	1	0.995	0.996	0.857
	Kurtosis	0.001	0	0.104	0.94	0.897	0.966	0.74
	GrayLevelVariance	0.001	0	0.011	1	0.995	0.837	0.857
	GrayLevelNonUniformityNormalized RunVariance	0.001	0	0	1	0	0.966	0
	LongRunEmphasis	0.001	SPIR	0.211	1	0.653	0.011	0.688
ngtdm	ShortRunHighGrayLevelEmphasis	0.008	0	1	0.92	1	0.659	0.494
	RunLengthNonUniformity	0.001	0	0.004	0.98	0.503	0.534	0.896
	LongRunHighGrayLevelEmphasis	0.001	0	0	1	1	0.981	1
	RunEntropy	0.001	0	0.008	1	1	1	0.87
	HighGrayLevelRunEmphasis	0.001	0	0	1	1	1	1
	GrayLevelVariance	0.001	0	0.006	1	1	1	0.883
	ZoneVariance	0.001	0	0.011	1	0.997	0.992	0.857
	SizeZoneNonUniformity	0.001	0	0	1	0.971	1	1
	GrayLevelNonUniformity	0.001	0	0	1	0.992	0.966	1
	LargeAreaEmphasis	0.001	0	0	1	1	0.962	1
	SmallAreaHighGrayLevelEmphasis	0.005	0	0	0.94	0.916	1	0.974
	LargeAreaLowGrayLevelEmphasis	0.001	0	0.002	1	0.989	0.939	0.922
	LargeAreaHighGrayLevelEmphasis	0.008	0	0.479	0.92	0.663	0.939	0.61
	HighGrayLevelZoneEmphasis	0.001	0	0	1	1	0.583	0.961
	SmallAreaEmphasis	0.001	0	0.008	1	1	1	0.87
	ZoneEntropy	0.001	0.002	0.659	1	0.561	0.989	0.429
	Complexity	0.001	0.002	0	1	1	0.485	1
	Busyness	0.001	0	0	0.98	0.989	1	1
		0.003	0	0.246	0.96	0.947	0.989	0.675
			0				0.807	
			0					
			0					
			0					
			0					
			0					
			0					
			0.009					
			0					
			0					
			0					
			0					
			0					
			0.316					

Table 2. (*Continued*).

Category	Feature	T1wI	P value	DELAY	TIWI	T2WI	ROCSPIR	DELAY
			0.001					
			0					
			0					
			0					
			0					
			0					
			0					
			0.006					
			0.174					
			0					
			0					
			0					
			0.001					
			0					
			0					
			0.142					
			0.766					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0					
			0.451					
			0					
			0					
			0.903					
			0					
			0					
			0.003					

A ROC curve of 105 features was performed to evaluate the ability of the features to distinguish hepatocellular carci noma from sclerosing nodules. This curve (AUC < 0.85) was abandoned in this study due to its limited discriminative ability. In the end, this study obtained a total of 68 characteristic ROC curves (**Figures 3**–**6**).

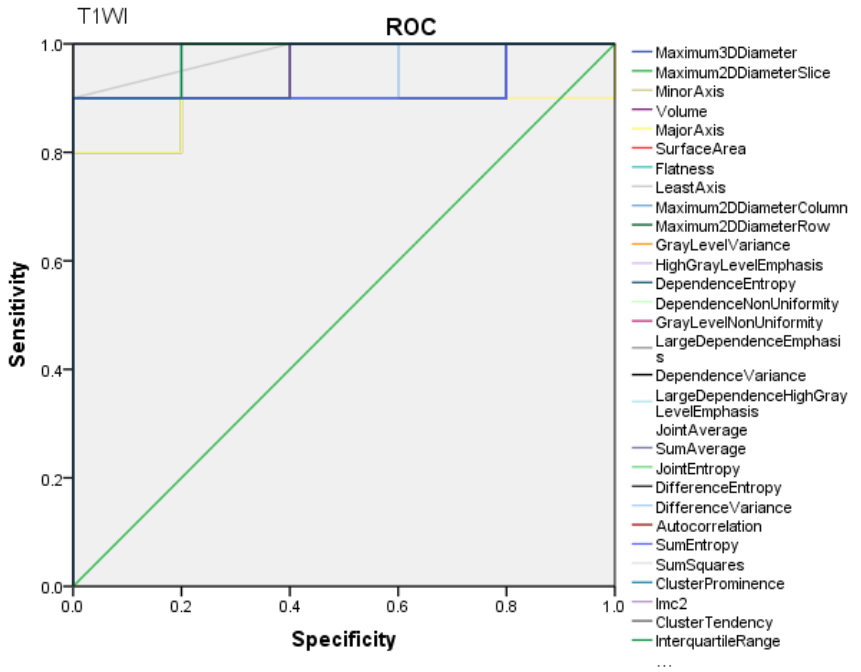


Figure 3. The ROC curves of T1WI.

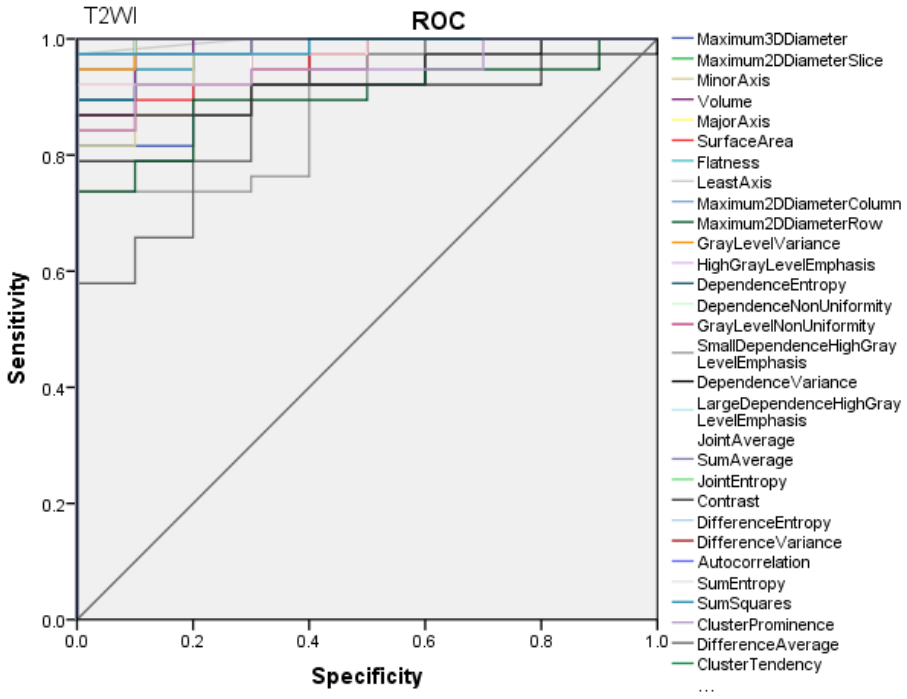


Figure 4. The ROC curves of T2WI.

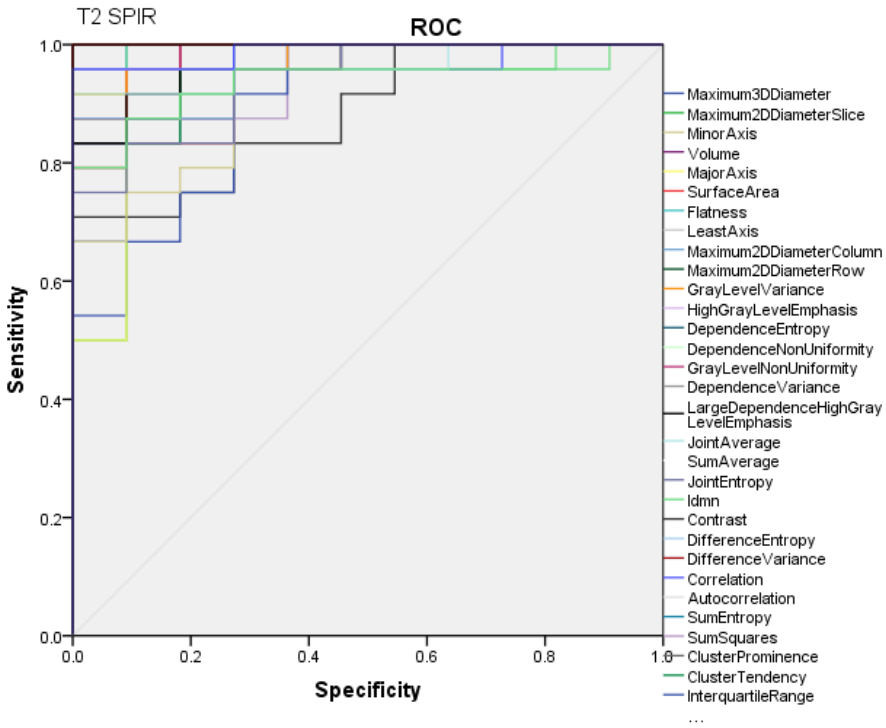


Figure 5. The ROC curves of SPIR.

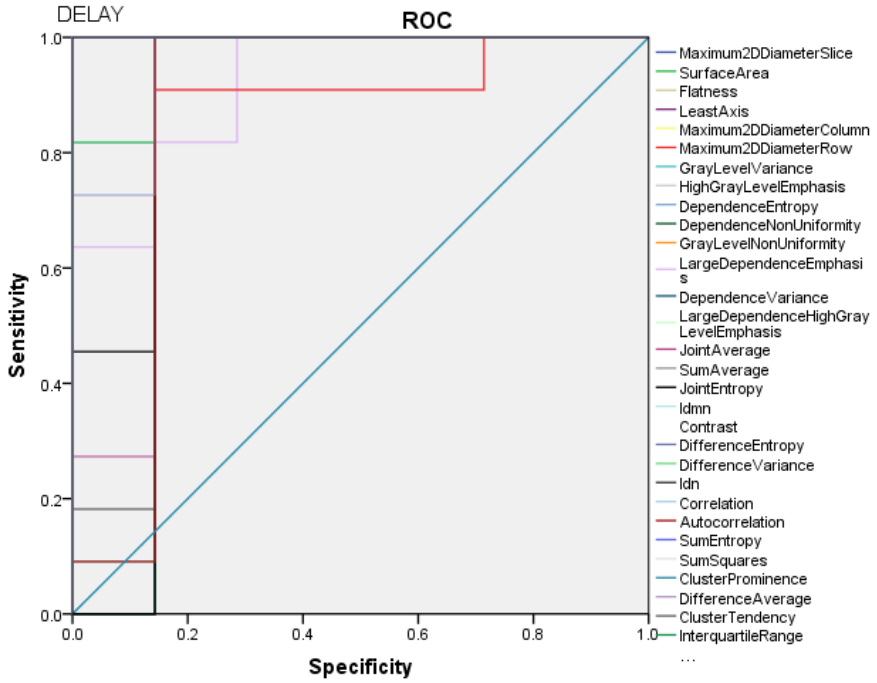


Figure 6. The ROC curves of DELAY.

5. Discussion

The results of this study show that there is a statistical difference between thera-diographic features extracted from hepa-tocellular carcinoma lesions and thera-diographic features extracted from cir-rhotic nodules. This maybe related to their different pathological tissue morphology. Carcinogenesis is a process in which non-

malignant liver cells gradually transform into liver cancer, which is a complex and multi-step process. For clinical practicability and research, this process is divided into several independ-ent steps: Cirrhotic nodules, dyplastic nodules, early liver cancer, and pro-gressed liver cancer [11]. This study selects the stage of cirrhotic nodules. Cirrhosis nodules,also called regenerative nodules related to liver cirrhosis,are countless clear circular areas of hard-enedparenchyma with scar tissue around them, with a diameter of 1–15 mm [11]. Cirrhotic nodules are generally considered benign because of its lacking histological features and normal phenotype [12]. But from a molecular perspective, many cirrhotic nodules are the clonal expansion of abnormal genomic cells, causing the macrophages in the cirrhotic nodules to develop abnormal proliferation characteristics [13]. So it will cause hyperplasia and nodules. A large number of previous studies have shown that the molecular changes of liver cells caused by abnormalities such as cell sig-nal transduction caused by chronic in-flammation begin in the early stage of tumor formation [14–16]. That is, several years or even decades before the onsetof liver cirrhosis, and with the development of fibrosis and cirrhosis Parallel development [17,18]. Studies have shown that the earliest molecular change in liver cancer is morphological silence, suggesting that chronically ill liver may contain cells with abnormal molecular but normal phenotypes, which will eventually develop into liver cancer [13,18,19]. Pathologically, early HCC is composed of small, well-differentiated neoplastic cells arranged in irregular but thintra-beculae or pseudogland [20], microscopically similar to highly hyperplastic nodules [21]. The tissues of advanced liver cancer lesions have the characteristics of mosaic structure, that is, there are multi-ple tumor nodules inside, and these nodules are separated by fibers, and there are areas of hemorrhage, necrosis, and occasional steatosis [22]. The subtle differences in histology between hepatocellu-lar carcinoma and cirrhotic nodules can be distinguished on MR.

The radiomics technology that has emerged in recent years refers to the high-throughput extraction of a large number of image features describing tumor characteristics, and the application of a large number of automated data retention methods to convert the image data of the region of interest into high-resolution imaging data. Feature space data sent [23,24]. Data analysis is a digital quantitative high-throughput analysis of a large amount of image data to obtain high-fidelity target information to comprehensively evaluate various phenotypes of tumors, including tissue morphology, cell molecular, genetic inher-itance and other levels. The core theo-retical basis is the radiomic model,which contains the biological or medical data information of the lesion, which can provide valuable information for the diagnosis, prognosis and prediction of the disease [25,26]. There is genetic heterogeneity among tumors of different patients, different tumor tissues of the same patient, or within the same tumor, and their genetic status will also vary from time to time. Based on the above advantages,some researchers have combined radiomics with medical images and applied them to tumor prediction, identification and prognosis. Imageomics has shown excellent performance in the diagnosis of lung cancer [27], stomach cacer [28], prostate cancer [29], and breast cancer [27]. Tsai et al. [30] reported that Texture features can be used to distinguish nasopharyngeal carcinoma from normal nasopharyngeal tissue, and the statistical difference in texture features between nasopharyngeal carcinoma and normal nasopharyngeal tissue maybe related to the loss

of stripe structure in normal nasopharyngeal tissue. and this finding had been confirmed on MRI images. Thawani et al. proposed that radiomics has played an important role in the diagnosis of lung cancer in recent years and will further provide more important information for monitoring and prognosis, and realize individualized treatment [31,32].

6. Conclusions

The results of this study show that MR is of great significance for the diagnosis of liver cancer, and imaging omics is of great value in the differentiation between benign and malignant lesions.However, the research method in this ar-ticle has limitations: (1) This article uses a single-center study with a small number of samples; (2) Lack of differentiation from patients without liver cancer; (3) Not combined with patient pathologi-cal smears; (4) Only one kind of imaging is used Methods, failed to compare the sensitivity and specificity of different imaging techniques to lesions. Our later research will try multi-center research to obtain a large number of samples based on more imaging methods to improve the accuracy of the results.

Author contributions: Conception and design, CM (Changdong Ma); administrative support, SY; provision of study materials or patients, CM (Changsheng Ma); collection, CM (Changsheng Ma); assembly of data, CM (Changdong Ma); data analysis, CM (Changsheng Ma); interpretation, CM (Changdong Ma); manuscript writing, CM (Changdong Ma); final approval of the manuscript, CM(Changdong Ma), CM (Changsheng Ma) and SY. All authors have read and agreed to the published version of the manuscript.

Ethical approval: The study was conducted in accordance with the Declaration of Helsinki. The protocol for this study was approved by the Institutional Review Committee of the Shandong First Medical University Affiliated Cancer Hospital. The ethics filing number is SDTHEC2020010008. As this is a retrospective study and sensitive information of all patients was hidden during the study process, so the need for informed consent was waived by the Institutional Review Committee of the Shandong First Medical University Affiliated Cancer Hospita (SDTHEC2020010008).

Availability of data and materials: The datasets used and analyzed during the current study available from the corresponding author on reasonable request (machangsheng_2000@126.com).

Funding: This study was supported by the National Nature Science Foundation of China (81974467), the Natural Science Foundation of Shandong Province (ZR2023MH166).Shandong Medical Association Clinical Research Fund-Qilu Special Project (YXH2022ZX02197), Shandong Traditional Chinese Medicine Technology Project (M-2022225).

Conflict of interest: The authors declare no conflict of interest.

Abbreviations

CT	Computed Tomography
MRI	Magnetic Resonance Imaging
DWI	Diffusion Weighted Imaging
T2WI	T2-Weighted Imaging
ROI	Region of Interest;
VOI	Volume of Interest;
ROC	Receiver Operating Characteristic;
AUC	Area Under the Curve
GLRLM	Gray Level Run Length Matrix;
GLCM	Gray Level Co-occurrence Matrix
GLSZM	Gray Level Size Zone Matrix
NGTDM	Neighborhood Gray-Tone Difference Matrix
GLD	Gray Level Dependence Matrix
CEUS	Contrast-enhanced Ultrasound
TR	Repetition Time
TE	Echo Time

Reference

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68(6): 394–424. doi:10.3322/caac.21492
2. McGlynn KA, London WT. Epidemiol-ogy and natural history of hepatocellu-lar car cinoma. Best Pract Res Clin Gas-troenterol. 2005;19(1): 3–23.
3. Theise ND, Curado MP, Franceschi S, et al. Hepatocellular carcinoma. In: Bosman FT, Carneiro F, Hruban RH, Theise ND. (editors). WHO Cassification of Tumours of the Di gestive System. IARC Publishing; 2010. pp. 205–216.
4. Choi JY, Lee JM, Sirlin CB. CT and MR imaging diagnosis and staging of hepatocellular carcinoma: part I. Devel-opment, growth, and spread: Key path-ologic and imaging aspects. Radiology. 2014; 272(3): 635–654.
5. Ronot M, Purcell Y, Vilgrain V. Hepato-cellular Carcinoma: Current Imaging Modalities for Diagnosis and Prognosis. Dig Dis Sci. 2019; 64(4): 934–950. doi:10.1007/s10620-019-05547-0
6. Forner A, Vilana R, Ayuso C, et al. Diag nosis of hepatic nodules 20 mm or smaller in cirrhosis: Prospective valida-tion of the noninvasive diagnostic cri-teria for hepato cellular carcinoma. Hepatology 2008; 47(1): 97–104.
7. Huang JY, Li JW, Lu Q, et al. Diagnostic Accuracy of CEUS LI-RADS for the Characterization of Liver Nodules 20 mm or Smaller in Patients at Risk for Hepatocellular Carcinoma. Radiology. 2020; 294(2): 329–339.
8. Chen X, Yang Z, Deng J. Use of 64-Slice Spiral CT Examinations for Hepatocel-lular Carcinoma (DR LU). J BUON. 2019; 24(4): 1435–1440
9. Di Martino M, De Filippis G, De Santis A, et al. Hepatocellular carcinoma in cir-rhotic patients: prospective comparison of US, CT and MR imaging. Eur Radiol. 2013; 23(4): 887–896. doi:10.1007/s00330-012-2691-z
10. Sun XJ, Quan XY, Huang FH, Xu YK. Quantitative evaluation of diffusion-weighted magnetic resonance imaging of focal hepatic lesions. World J Gastro-enterol. 2005; 11(41): 6535–6537. doi:10.3748/wjg.v11.i41.6535
11. International Working Party. Terminol-ogy of nodular hepatocellular lesions. Hepatol ogy. 1995; 22(3): 983–993.
12. Park YN, Kim MJ. Hepatocarcinogenesis: imaging-pathologic correlation. Ab-dom Im aging 2011; 36(3): 232–243.
13. Aihara T, Noguchi S, Sasaki Y, Nakano H, Imaoka S. Clonal analysis of regen-erative nodules in hepatitis C virus-in-duced liver cirrhosis. Gastroenterology. 1994; 107(6): 1805– 1811.
14. Trevisani F, Cantarini MC, Wands JR, Bernardi M. Recent advances in the nat ural history of hepatocellular carcinoma. Carcinogenesis. 2008; 29(7): 1299–1305.
15. Brody RI, Theise ND. An inflammatory proposal for hepatocarcinogenesis. Hepa tology 2012; 56(1): 382–384.

16. Thorgeirsson SS, Grisham JW. Molecu-lar pathogenesis of human hepatocel-lular car cinoma. Nat Genet, 2002; 31(4): 339–346.
17. Theise ND. Macroregenerative (dys-plastic) nodules and hepatocarcino-genesis: theo retical and clinical consid-erations. Semin Liver Dis .1995;1 5(4): 360–371.
18. Aravalli RN, Cressman EN, Steer CJ. Cel lular and molecular mechanisms of hepato cellular carcinoma: An update. Arch Toxicol. 2013; 87(2): 227–247.
19. Sun M, Eshleman JR, Ferrell LD, et al. An early lesion in hepatic carcinogenesis: loss of heterozygosity in human cir-rhotic livers and dysplastic nodules at the 1p36-p34 region. Hepatology .2001; 33(6): 1415–1424.
20. Park YN. Update on precursor and early lesions of hepatocellular carcinomas. Arch Pathol Lab Med. 2011; 135(6): 704–715.
21. Roskams T, Kojiro M. Pathology of early hepatocellular carcinoma: conventional and molecular diagnosis. Semin Liver Dis 2010; 30(1): 17–25.
22. Stevens WR, Gulino SP, Batts KP, et al. Mosaic pattern of hepatocellular carcinoma: histologic basis for a characteristic CT appearance. J Com put Assist Tomogr. 1996; 20(3): 337–342.
23. Lambin P, Rios-Velazquezet E, Leijenaaral R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012; 48: 441–446.
24. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magn. Reson. Im-aging. 2012; 30, 1234–1248.
25. Haase AT, Henry K, Zupancic M, et al. Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science. 1996; 274, 985–989.
26. Schoolman H, Bernstein L. Com-puter use in diagnosis, prognosis, and therapy. Science. 1978; 200: 926–931.
27. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. Strahlenther Onkol. 2020; 196(10): 879–887. doi:10.1007/s00066-020-01625-9
28. Jiang Y, Chen C, Xie J, et al. Radiomics signature of computed tomography imaging for prediction of survival and chemotherapeutic benefits in gastric cancer. EBioMedicine. 2018; 36: 171–182. doi:10.1016/j.ebiom.2018.09.007
29. Smith CP, Czarniecki M, Mehralivand S, et al. Radiomics and radiogenomics of prostate cancer. Abdom Radiol (NY). 2019; 44(6): 2021–2029. doi:10.1007/s00261-018-1660-7
30. Tsai A, Buch K, Fujita A, et al. Using CT texture analysis to differentiate between naso-pharyngeal carcinoma and age-matched adenoid controls. Eur J Radiol. 2018; 108: 208–14.
31. Thawani R, McLane M, Beig N, et al. Radiomics and radiogenomics in lung cancer: A review for the clinician. Lung Cancer. 2018; 115: 34–41. doi:10.1016/j.lungcan.2017.10.015
32. Wei K, Su H, Zhou G,et al. Potential application of radiomics for differentiating solitary pul-monary nodules,OMICS J Radiol. 2016; 5(2): 1000218

Imaging and Radiation Research 2024, 7(1), 6257.
https://doi.org/10.24294/irr6257

Article

Classification of X-ray images and model evaluation

Aya Naser[1], Şafak Bera Şafak[1], Emrah Utkutağ[1], Simge İnci Sin[1], Sena Sude Taşkin[1], İrem Koca[1], Refika Sultan Doğan[1,2,*]

[1] Department of Bioengineering, Faculty of Life and Natural Sciences, Abdullah Gül University, Kayseri 38080, Turkey
[2] Biomedical Instrumentation and Signal Analysis Laboratory, Faculty of Engineering, Abdullah Gül University, Kayseri 38080, Turkey
*** Corresponding author:** Refika Sultan Doğan, refikasultan.dogan@agu.edu.tr

CITATION

Naser A, Şafak SB, Utkutağ E, et al. Classification of X-ray images and model evaluation. Imaging and Radiation Research. 2024; 7(1): 6257. https://doi.org/10.24294/irr6257

ARTICLE INFO

Received: 7 May 2024
Accepted: 13 June 2024
Available online: 21 November 2024

Abstract: Inflammation of the lungs, called pneumonia, is a disease characterized by inflammation of the air sacs that interfere with the exchange of oxygen and carbon dioxide. It is caused by a variety of infectious organisms, including viruses, bacteria, fungus, and parasites. Pneumonia is more common in people who have pre-existing lung diseases or compromised immune systems, and it primarily affects small children and the elderly. Diagnosis of pneumonia can be difficult, especially when relying on medical imaging, because symptoms may not be immediately apparent. Convolutional neural networks (CNNs) have recently shown potential in medical imaging applications. A CNN-based deep learning model is being built as part of ongoing research to aid in the detection of pneumonia using chest X-ray images. The dataset used for training and evaluation includes images of people with normal lung conditions as well as photos of people with pneumonia. Various preprocessing procedures, such as data augmentation, normalization, and scaling, were used to improve the accuracy of pneumonia diagnosis and extract significant features. In this study, a framework for deep learning with four pre-trained CNN models—InceptionNet, ResNet, VGG16, and DenseNet—was used. To take use of its key advantages, transfer learning utilizing DenseNet was used. During training, the loss function was minimized using the Adam optimizer. The suggested approach seeks to improve early diagnosis and enable fast intervention for pneumonia cases by leveraging the advantages of several CNN models. The outcomes show that CNN-based deep learning models may successfully diagnose pneumonia in chest X-ray pictures.

Keywords: convolutional neural networks; image classification; image processing; medical imaging; artificial intelligence

1. Introduction

Pneumonia is an infection-related inflammation of the lungs' air sacs (alveoli). Alveoli is found at the ends of the respiratory bronchioles, allowing the exchange of oxygen and carbon dioxide gas. It may be referred to as bronchopneumonia if the airways are also affected. Pneumonia occurs when these air sacs are filled with fluid or pus, hindering the gas exchange process, resulting in difficulty breathing and a cough reflex. It can affect the lung in one or more sites (sometimes known as "double" or "multilobar" pneumonia). pneumonia can be caused by a variety of factors, the majority of which are infectious [1].

Pneumonia is typically caused by virus or bacterial infection from the environment or from another person. Infection can be spread from person to person by direct touch (typically through the hands) or through inhaling droplets in the atmosphere from coughing or sneezing. Secondary infection from bacteria like Staphylococcus aureus can occasionally occur in a person who has a viral illness, such as the influenza virus, while they are ill. Pneumonia can also be caused by a parasite,

fungus, or yeast. Aspiration pneumonia is brought on by a foreign substance entering the lungs through the throat. Typically, this material is food or vomit, which causes irritation to the airways and lung tissue and raises the risk of bacterial infection [1].

Pneumonia can occur at any age. However, it seems to affect small children and the elderly more frequently. Pneumonia can be more serious for some people due to pre-existing lung conditions, poor nutrition, swallowing issues, other chronic health issues, or immune system issues. Pneumonia is more likely to occur among smokers and those who are around tobacco smoke. People who have not had the annual influenza vaccination or the pneumococcal vaccines Prevnar13® and/or Pneumovax®23 are also at a higher risk of developing lung infections [2].

People who have pneumonia frequently experience coughs, fever or chills, respiratory problems, low energy, and poor appetite. Chest pain, nausea, and/or diarrhea can all occur frequently. Without a cough or fever, pneumonia is possible. Symptoms may appear suddenly or develop gradually over time. An individual who has a viral upper respiratory illness (cold) may occasionally experience a new fever and deterioration, which indicates the beginning of the secondary bacterial infection [2].

The medical professional will take the symptoms into account and do a physical assessment. Pneumonia can cause the noises in the lung to be diminished or irregular. Blood tests may be performed to check the white blood count and other measurements that may be off related to a disease. A chest x-ray is frequently taken in order to identify the site or regions affected by pneumonia. Sometimes a CT scan—often referred to as a "cat" scan—is performed for more precise computerized x-rays [3]. Sputum, also known as phlegm or mucus, is excreted during coughing, and may be tested and cultured to determine the presence of any germs or viruses. More frequently, tests are performed on patients who are ill enough to be hospitalized for the most likely viruses and bacteria. A procedure known as flexible bronchoscopy may be used to extract a sample of mucus from the lung through the airways if a patient is not improving, has a serious infection, or is at a high risk of developing a rare infection [3]. It can be difficult to determine what type of infection (for example, which bacterium) is causing pneumonia. This could be a result of the tests being insufficiently precise or because you might have undergone treatment prior to the testing. However, the healthcare provider will work with the patient to choose a course of action based on the most likely cause determined by the patient's information, the types of infections that are prevalent in the patient's community, and the types of infections that the patient may be more susceptible to if they already have a health issue [3]. The prognosis for pneumonia and the seriousness of the patient's condition both influence treatment. Antibiotics that are efficient against the most likely microorganisms causing the infection are typically given. The patient may require medications to address more resistant germs if the pneumonia occurred while they were a patient in a hospital or another healthcare facility such as a nursing home [3]. Because the symptoms of the illness are not readily apparent on CT or X-ray scans, pneumonia objectively and automatically detecting poses a significant issue in medical imaging. Chest X-rays (CXR) or computed tomography (CT) scans are frequently used to detect pneumonia; the former is the most used method because it is more affordable and widely available worldwide. Due to its great speed and objective, reproducible judgment, the computer can assist the human expert in making the

diagnosis of pneumonia because the symptoms of pneumonia in X-ray images are not always obvious or readable to the human eye [4]. For the purposes of computer vision, researchers have suggested various CNN-based deep networks for image classification, picture segmentation, object recognition, and localization. CNNs have proven to be extremely effective and successful at resolving medical issues as well, including the diagnosis of Alzheimer's disease, the classification of skin lesions, the identification of breast cancer, and the segmentation of brain tumors [5].

Machine learning and deep learning methods have recently been developed in numerical computing for medical picture analysis. Since they offer great accuracy and amazing outcomes when compared to other models, convolutional neural networks (CNNs) are the most favored and well-liked deep learning models with superior accomplishments in the medical imaging sector. On the basis of the CNNs, numerous research using various methods to perform chest X-rays was undertaken to identify pneumonia. A CNN model, for instance, was proposed by Stephen et al. and trained to categorize pneumonia using chest X-rays. The proposed model's accuracy, according to the authors, is 95.31% [6]. Convolutional neural networks (CNNs) were used in picture classification tasks, and their use of Deep Learning (DL) models proved their potential for doing so. This feature-extraction approach necessitates transfer learning techniques, in which pre-trained CNN models first learn the generic features on massive datasets like ImageNet, then transfer those features to the desired job. The process of important feature extraction is greatly facilitated by the availability of pre-trained CNN models like AlexNet, VGGNet, Xception, ResNet, and DenseNet. Additionally, the classification of photos when using highly-rich extracted features performs better. The datasets used, which include 112,120 anterior chest X-ray pictures from 30,085 patients, are also freely accessible on the Kaggle platform [7].

Since deep CNN models like ResNet, Xception, or DenseNet have millions of trainable parameters, training them from the start takes a large amount of data because the model wouldn't be sufficiently generalized with a small dataset. A TL approach can be used to reuse these models with their pre-trained weights. A pre-trained CNN model is reused in TL, a helpful machine learning technique, to use its weights as initialization for a new CNN model that will be used for a different task. The two main approaches to use the TL from a model are to either reuse the model to do Fine-Tuning (FT) or to reuse the model as a feature extractor and use an entirely new classifier. FT is a strategy that modifies the new Fully Connected (FC) layers of the classifier as well as particular layers of the CNN, such as convolutional layers, somewhat [8].

2. Materials and method

2.1. Dataset

A dataset of chest X-ray images labeled as Pneumonia and normal was used for detection with CNNs (**Figure 1** and **2**). The "Chest X-Ray Images (Pneumonia)" dataset was utilized in the code. The dataset consists of a series of chest X-ray images divided into 4273 "PNEUMONIA" images and 1583 "NORMAL" images categories in total. X-ray images of patients with pneumonia are included in the "PNEUMONIA" category, whereas images of healthy people without pneumonia are included in the "NORMAL" category.

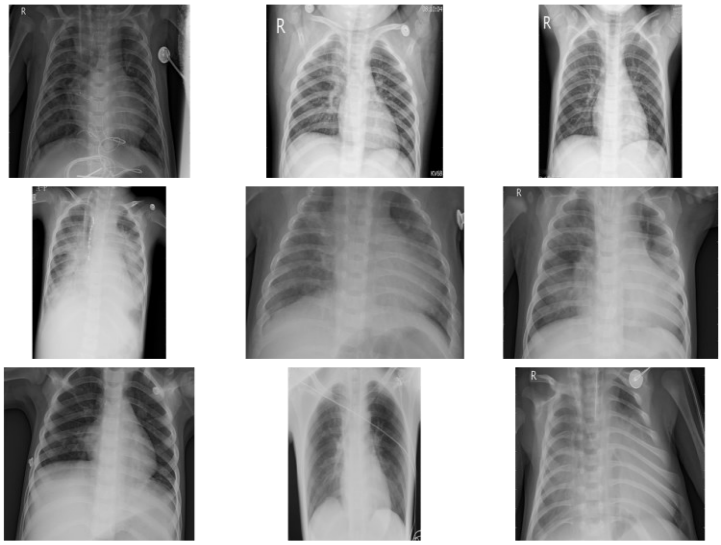


Figure 1. Chest X-Ray images of pneumonia patients.

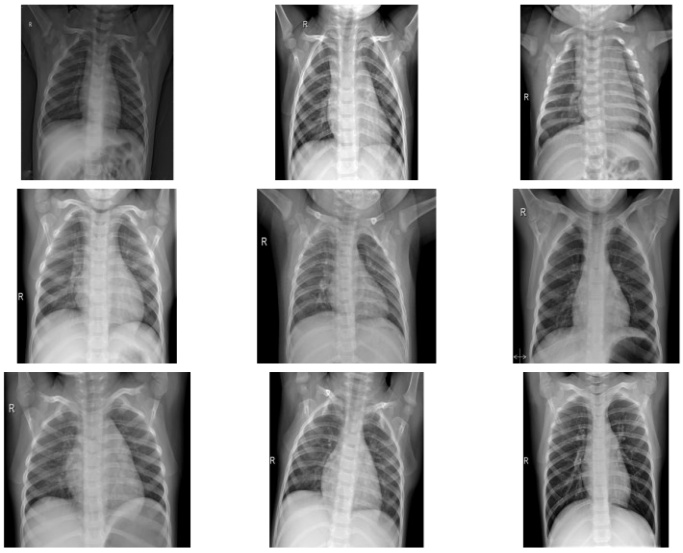


Figure 2. Chest X-Ray images of normal (non-pneumonia) patients.

2.1.1. Data pre-processing

The necessary pre-processing steps were applied on the X-ray images as follows: Before training, the images are modified for better training of a convolutional neural network. The Keras ImageDataGenerator function was used to prepare X-ray images for training a convolutional neural network (CNN). Image data augmentation is a method for enhancing the variety and variability of training data for deep learning models. The ImageDataGenerator provides random variations to the images during training by changing transformation parameters including rotation range, width shift range, shear range, and zoom range. Furthermore, the choice of samplewise centering and standardization helps in separately normalizing the pixel intensity of each image. However, this ImageDataGenerator combines these techniques to generate enhanced image batches that may be used to train deep learning models, enhancing generalization and robustness by offering a more diverse and variable training set.

2.1.2. Data split

The dataset was divided into three subsets: Train, validation, and test sets (**Table 1**).

Table 1. The number of input images with normal and pneumonia chest X-ray images.

	Normal	Pneumonia
Train	1341	3875
Test	234	390
Val	8	8

2.1.3. Data augmentation (normalization and resizing)

To standardize the input distribution and facilitate model training, the pixel values in each set of images were fixed in a range from 0 to 1, and the images were resized to a consistent size. This analysis provided valuable information about the intensity and spread of pixel values across the dataset, aiding our understanding of the data characteristics.

2.2. Channel conversion

Our chosen pre-trained neural network requires three-channel input. Hence, we utilized the generator to convert the single-channel grayscale X-ray images into a three-channel format. This conversion was achieved by replicating the pixel values across all three channels.

2.2.1. Convolutional neural networks (CNNs) model building

CNNs are deep learning architectures that excel at extracting features from images before classifying them. By employing convolution filters with different dimensions or values, various features can be extracted from the images. Features are detected using ReLu activation at each pixel and enhanced with MaxPool layers. The stride parameter determines the distance between each filter, while the padding parameter determines whether the mesh should consider boundary pixels. Zero padding applied to the neural network to provide information from image borders. The outputs from these operations were combined and passed through Dense layers. A sigmoid activation function was used to determine the final layer of the network, the class to which the image belongs.

2.2.2. CNNs model training

Training images were fed to CNN. Used a weighted loss function, such as binary cross-entropy, to balance the contribution of several classes to the loss calculation.

A sequential CNN framework for identifying pneumonia (pna) in clinical images is shown in **Figure 3**. Four pre-trained neural network (CNN) models are included in the framework: InceptionNet, ResNet, VGG16, and DenseNet. These pre-trained models are effective tools for feature extraction and picture segmentation tasks since they have been refined and optimized on huge datasets. The framework can benefit from the information and representation acquired from various datasets thanks to the use of pre-trained CNN models. The system gains from the capacity to extract useful

information from input photos and generate precise predictions based on these retrieved features by adding these models. A CNN architecture called InceptionNet introduces the idea of inception modules, which successfully capture characteristics at various spatial scales. ResNet, on the other hand, makes use of residual connections to help deep network training without experiencing vanishing gradient problems. A well-liked CNN architecture noted for its ease of use and effectiveness in image classification tasks is VGG16. Finally, DenseNet uses densely interconnected blocks to enhance information flow across layers, making layer use simpler and lowering the possibility of overpopulation. The inclusion of these four pre-trained CNN models in **Figure 3** shows that the design takes advantage of their unique strengths and different design methods to improve the accuracy and reliability of pneumonia diagnosis. By combining the capabilities of these models, the design aims to capture a wide range of image features and provide powerful predictions for the detection of pneumonia, thus facilitating early diagnosis and timely medical intervention.

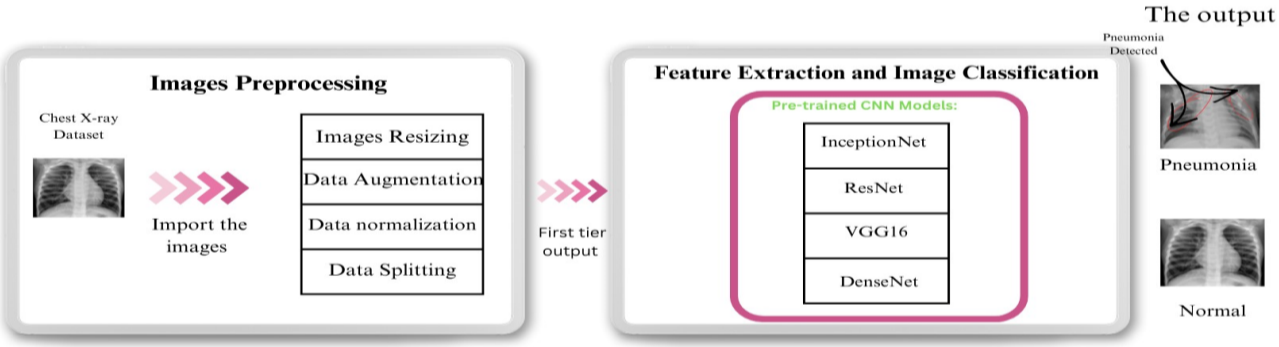


Figure 3. The deep learning framework that is suggested for diagnosing pna.

$$L^{W}_{cross_{entropy}}(x) = -\left(w_p y \log\left(f(x)\right) + w_n (1-y) \log\left(1-f(x)\right)\right) \quad (1)$$

Loss function formula explained in Equation (1). Class weights were applied for class 0 and class 1, which were applied to tolerate the imbalance condition. Weight for class 0: 0.74, weight for class 1: 0.26 were found. The Adam optimizer was used to update the model weights and minimize the loss value. Transfer Learning was employed using DenseNet, a convolutional neural network architecture. DenseNet is characterized by dense connections between layers, where each layer is connected to all the subsequent layers in the network. This connectivity pattern facilitates the flow of gradients throughout the network, enabling efficient feature propagation.

2.2.3. Transfer learning models training

InceptionNet

Google created the convolutional neural network (CNN) architecture known as InceptionNet to outperform earlier CNNs. It is well-known for using inception modules, which are layers' building blocks that learn a mix of local and global features from the input data. There are 22 layers in the InceptionNet architecture, including fully connected, pooling, and convolutional layers. The use of "Inception modules", parallel convolutional blocks with various filter sizes (1 × 1, 3 × 3, and 5 × 5) and pooling operations, is one of its key innovations. The network can learn spatial and

temporal features from the input data using these modules, which are made up of smaller convolutional and pooling layers that are combined. Other deep convolutional neural networks can take longer to train than InceptionNet because of the way it was designed. It serves as the foundation for well-known neural network architectures like Inception-v4 and Inception-ResNet and has been used in image classification, object detection, and face recognition. The InceptionNet model has roughly 5 million parameters total. But over time, a few variations, and more advanced versions of InceptionNet, including InceptionV2, InceptionV3, InceptionV4, and Inception-ResNet, each with a different number of parameters, were developed. Depending on the depth and complexity of the architecture, these variants have anywhere between tens of millions and hundreds of millions of parameters. The complexity of the computations is typically increased in the deeper versions in exchange for performance gains [9].

Conventional convolutional neural networks often use convolutional and pooling layers to extract features from the input data. This problem is resolved by Inception blocks' modular nature, which allows the network to learn various feature maps at varied scales. These feature maps are then concatenated to produce a more comprehensive representation of the input data. To help with tasks like picture categorization, this enables the network to collect a variety of features, both high-level and low-level. Using inception blocks allows the InceptionNet architecture to learn a larger range of features from the input data, which improves the network's performance on tasks like picture categorization. The Inception network is made up of convolutional design configurations in the form of recurring patterns known as Inception modules [10].

- Input layer
- 1 × 1 convolution layer
- 3 × 3 convolution layer
- 5 × 5 convolution layer
- Max pooling layer
- Concatenation layer

The Inception modules are a key part of the InceptionNet convolutional neural network architecture. These built-in blocks, known as layers, are designed to extract a combination of local and global properties from the incoming data. By combining smaller convolutional and pooling layers from inception modules, the network may extract spatial and temporal properties from the input data. The objective of the inception module is to learn many feature maps at different scales, then integrate them to produce a more comprehensive representation of the input data. The network may consequently gather a wide range of low-level and high-level information that can be useful for tasks like picture categorization. Depending on the desired level of complexity and the volume of input data, inception modules can be added to the network at different points. They can also be altered by altering the convolutional and pooling layers' size and number, as well as the nonlinear activation function's type.

The necessary TensorFlow libraries and classes are imported. The input shape and number of classes for images are specified, including pneumonia and normal. To customize the pre-trained InceptionNet model for the classification task, additional custom layers are loaded. The model is put together using the Adam optimizer and

categorical cross-entropy loss after the pre-trained layers have been filled in. To perform real-time data augmentation during training, ImageDataGenerator is set up. The heap size and training period count are then set, along with the indexes for the train, validation, and test sets of data. Train loads and preprocesses validation and test data using flow_from_directory. Tags are categorically encoded, and images are resized. The model is trained by passing the train and validation data, the number of steps per period, and the validation steps through the fit function. Using the evaluate and print test loss and accuracy option, the model is evaluated on the test data after the training.

ResNet

Using the layer inputs as a guide, the weight layers of a residual neural network develop residual functions. Identifier mappings are carried out by skip connections in a residual network, which is added to the layer outputs. The strategy behind this network is to let the network fit the residual mapping rather than have layers learn the underlying mapping [11]. Thus, let the network fit instead of using, say, the initial mapping of $H(x)$,

$$\mathrm{F}(x) = H(x) - x \; which \; gives \; H(x) = F(x) + x$$

In this section, we'll go through how to classify images in Keras using ResNet50. Import libraries: Import Keras and other necessary libraries. Install ResNet50 and use Keras to initialize the ResNet50 model. The intended shape of the input photos is specified by the input shape argument. Set include top=False to prevent ResNet50's fully linked layers from being included in the model. For certain classification jobs, this enables customization. Use pre-trained weights: To start the model with pre-trained weights from the ImageNet dataset, set weights = 'imagenet'. These weights offer a place to start for accurate categorization.

An effective tool for picture categorization in Keras is ResNet50. By loading the model, setting it up, and using weights that have already been trained, accurate results can be obtained with minimal work. The model must be assembled, and the dataset must be ready, for the implementation to be complete. Using the Keras API of TensorFlow, additional code constructs a neural network model based on the ResNet architecture. The model is made up of numerous fully connected layers that are placed on top of a pre-trained ResNet basic model. The model is intended to solve a binary classification problem in which one of two classes is to be determined for each input. For enhancing model performance and avoiding overfitting, the design contains additional layers for feature extraction and editing. Accuracy, precision, and recall are just a few of the criteria that have been established to gauge how well the model is working. These metrics shed light on several facets of the categorization performance of the model. Overall, the code builds a binary classification model based on ResNet, compiles it using Adam optimizer and binary cross-entropy loss, and sets evaluation metrics to track the model's performance during training and evaluation.

VGG16

The VGG model, commonly known as VGGNet, is referred to as VGG16. It is a 16-layer convolutional neural network (CNN) model. When using numerous smaller layers rather than a single large layer, the decision functions are improved, and the

network can converge more quickly because there are more non-linear activation layers present [12]. In the top five tests, the model performs 92.7% accurately in ImageNet, a dataset of over 14 million images divided into 1000 classes [8].

VGG16 operating principle is based on the idea that stacking many smaller layers increases the network's decision-making power. The model can capture and learn complicated characteristics from input photos by employing multiple convolutional layers with non-linear activation functions. The input image for VGG16 is typically 224 × 224 pixels in size and is processed through several convolutional layers. Each convolutional layer extracts various features at various degrees of abstraction from the input by applying a set of learnable filters to the input [11]. These filters have been trained to recognize forms, edges, and other visual patterns. A non-linear activation function, such as the Rectified Linear Unit (ReLU), is used after each convolutional layer. By introducing non-linearity, the activation function enables the model to learn intricate connections between the retrieved data. The max-pooling layers in VGG16 also help to decrease the spatial dimensions of the feature maps by choosing the maximum value within a constrained area. Using methods like backpropagation and gradient descent, VGG16 learns to modify the weights and biases of its layers during training. To reduce the discrepancy between the model's anticipated outputs and the ground truth labels provided in the training data, the model is trained.

DenseNet

Each layer's features and gradients are strengthened by DenseNet by using the top classifier to oversee the other layers through feature connection. The efficiency of features from each hidden layer is less enhanced or verified by the top classifier, which is more likely to evaluate the effectiveness of the sum of input features for the final layer [13].

It uses convolutional layers, pooling, and dense blocks to obtain significant representations from input images. The input images' sizes are specified by the input shape option, and the final classification layers are disregarded if include top is set to False. In addition, pre-trained weights from the ImageNet dataset are loaded by setting weights to "imagenet", assisting with model initialization. After the convolutional layers, the pooling operation is determined by the pooling parameter. The base_model.summary() method displays the architecture's layers and parameter counts. The efficacy of DenseNet in many computer vision applications is mostly a result of its dense connections and configurable parameters [13].

2.3. Performance metrics

To measure performance, it is necessary to evaluate the trained model using the validation dataset. Accuracy, precision, recall, and F1-score are evaluation metrics for binary classification issues. The effectiveness of the classification models can be assessed in several different ways [14]. To assess the performance for categorizing colon polyps, we employed accuracy (Equation (2)), precision (Equation (3)), recall (Equation (4)), and f-measure (Equation (5)) measures.

$$Accuracy = \frac{(t_p + t_n)}{(t_p + f_p + f_n + t_n)} \quad (2)$$

$$Precision = \frac{t_p}{(t_p + f_p)} \tag{3}$$

$$Recall = \frac{t_p}{(t_p + f_n)} \tag{4}$$

$$f - measure = \frac{2}{(\frac{1}{Precision} + \frac{1}{Recall})} \tag{5}$$

All measures distinguish the correct classification of labels within different classes. Recall is a function of its correctly classified examples and its misclassified examples.

2.4. Model fine-tuning

Hyperparameters can be tuned, architecture changed, or other techniques such as normalization can be used to improve the performance of the model. Architectures such as VGG16, ResNet, InceptionNet can be used (**Figure 4**).

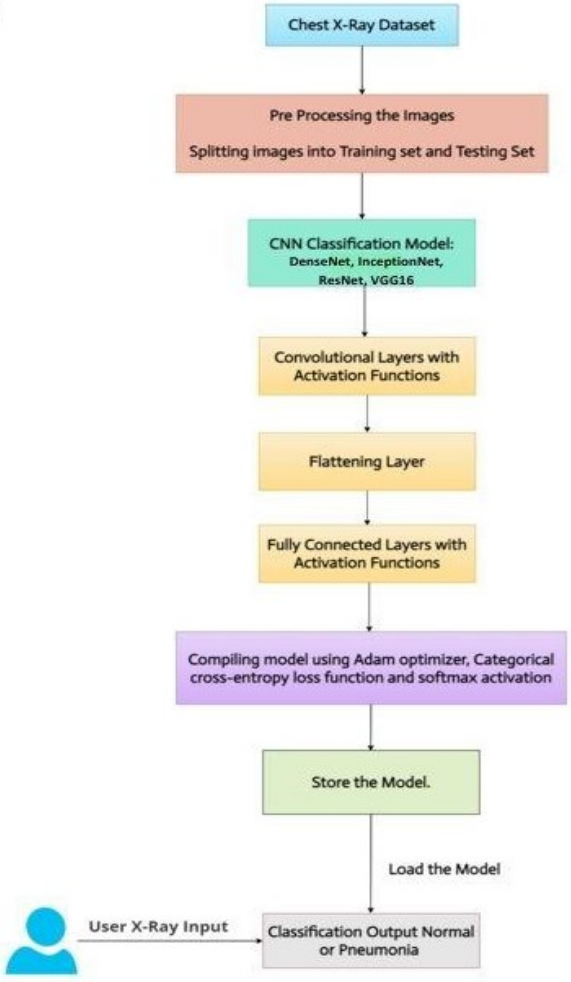


Figure 4. Model of pneumonia detection using convolutional neural networks from chest X-ray images.

3. Results and discussion

Table 2 shows a summary of the distribution of X-ray images for the training, testing, and validation datasets. The table has two categories: "Pneumonia" and "normal lung", which represent the presence or absence of pneumonia on X-ray images.

Table 2. Length of the input files with normal and pneumonia chest X-ray images.

	Normal	**Pneumonia**
Train	1341	3875
Test	234	390
Val	8	8

Table 3. Breakdown of the numbers in each dataset.

	Train	**Test**	**Validation (Val)**
Pneumonia	3875	390	8
Normal Lungs	1341	234	8

In the training list, there are 3875 X-ray images labeled as pneumonia and 1341 X-ray images labeled as normal lung. These images may be used to train a machine learning model to distinguish between pneumonia and normal lung X-ray images. The test dataset contains 390 pneumonia X-ray images and 234 normal lung X-ray images (**Table 3**). This dataset is often used to test the performance of a model trained on unobserved data. Validation data, with only 8 images for each class, is usually used to adjust or validate the model after the training and testing phases. The small size of the validation dataset indicates that only a small subset of the data is used for final model testing or hyperparameter tuning. These numbers provide an overview of how X-ray images are distributed in different datasets for the task of diagnosing pneumonia.

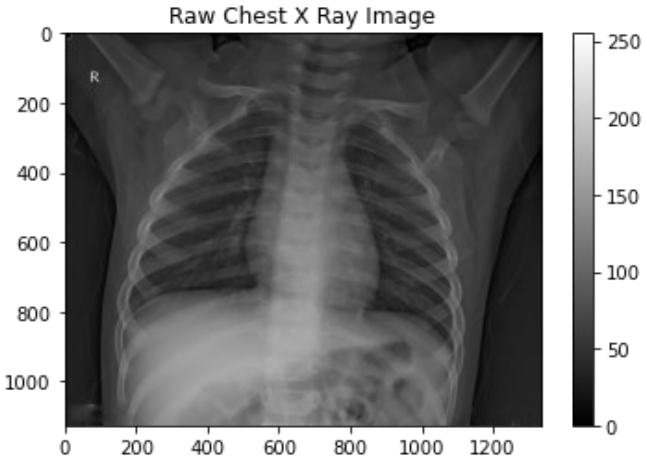


Figure 5. Pixel distribution and values.

According to **Figure 5**, the X-ray images in this study have a resolution of 1128 pixels and a height of 1336 pixels, with one color channel The pixel density ranges

from a minimum value of 0.00000 to a value of a maximum of 255.00000, representing the darkest and lightest values of the pixel, respectively. Graph 1 offers a pixel density distribution plot, which indicates the frequency of different pixel intensities inside the snap shots. The graph offers records for the X-ray pix. The measured pixel density is calculated as 73.2978, this means that a mean pixel density. However, the same old deviation of 38.1653 shows the distinction in pixel depth of many of the photographs. This statistical information is important for know-how the houses of X-ray pix and can help to develop picture processing techniques for medical applications.

3.1. Image preprocessing

In the context of image processing using the Python programming language, a set of parameters is used to add and manipulate images. These parameters include rotation range, width_shift_range, shear range, zoom range, and samplewise_center. The rotation range parameter allows random rotation of images up to 20 degrees, introducing contrast and improving the diversity of the dataset. The width_shift_range parameter enables the shift of image pixels by 10% of the image width, to produce a small change that simulates a different view or position. The shear range parameter introduces shear changes to images, enabling image shapes to be distorted by up to 10% of the image width. The zoom range parameter allows one to randomly zoom in or out of images by up to 10%, providing an additional level of flexibility and scale. Finally, the samplewise_center parameter sets the pixel values of each image by normalizing the images and subtracting the median value of the dataset. Together these parameters contribute to data optimization and manipulation techniques, facilitating the training and evaluation of machine learning models by expanding the data and improving its diversity.

3.2. Separate generator for valid and test sets

The reason why we cannot use a single generator for validation and analysis of the training data is because of the general operation performed by the generator. In training, the generator normalizes each image using batch statistics. This means that the mean and standard deviation used for normalization are calculated based on the images in the batch being processed. However, when it comes to validation and testing, we want to simulate a real situation where we deal with images individually and not in groups. In this case, the model should not have any knowledge about the test data beforehand. If we were to use a single generator with “batch normalization” for validation and analysis, it would provide information about the test data implicitly by allowing it to compute several batches. To avoid this issue, we need to use a separate generator for validation and test data. This generator needs to simplify the incoming check information and the usage of statistics taken from the training middle. Using popular facts used while training, we ensure that the version obtains regular and independent information during validation and assessment. This approach helps to hold the integrity of the assessment procedure, because the model is tested on random statistics without prior know-how or advantage received from the test institution. It permits us to accurately look at the version’s performance and determine how properly it generalizes to actual-international information. By following this method, we

expand a robust and independent evaluation framework for the overall performance of our version.

The photo in **Figure 6** is an X-ray image, that is represented in a virtual layout. The dimensions of this X-ray picture are precise as 180 pixels in width and 180 pixels in top. This method consists of a grid of one hundred eighty pixels horizontally and one hundred eighty pixels vertically. Additionally, it is cited that the X-ray picture consists of an unmarried color channel. In the context of grayscale pix, inclusive of X-rays, an unmarried coloration channel represents the depth or brightness values of each pixel. This indicates that the photograph is represented in shades of gray instead of containing coloration facts. By having these dimensions and a single-color channel, the X-ray photo in **Figure 7** can be processed and analyzed using numerous laptop imaginative and prescient techniques. Understanding the characteristics of the photo, which include its size and color channel, is critical for in addition evaluation, interpretation, and capability application of photograph processing algorithms or device mastering models within the scientific area. The pixel density graph in **Figure 8** shows the distribution of pixel intensity in the X-ray image shown in 6. The graph shows the frequency or number of pixels at different intensity levels. The observed statistics provide some details of the image, where the maximum and minimum pixel values of 2.5969 and −2.4856 respectively indicate the brightest and darkest pixels. With a total value of 0.0000, the overall strength appears to be around the neutral level. Furthermore, a standard deviation of 1.0000 suggests a wide range of pixel intensities throughout the image. These statistical studies contribute to a better understanding of image characteristics and can facilitate image processing techniques and medical applications.

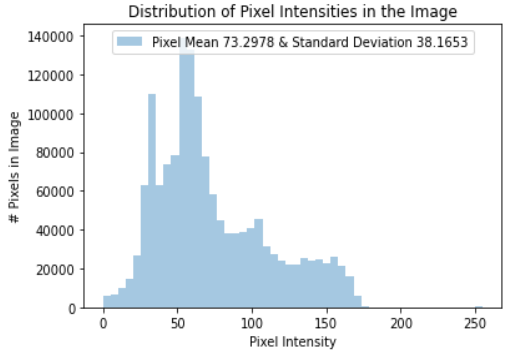


Figure 6. Pixel intensity distribution graph.

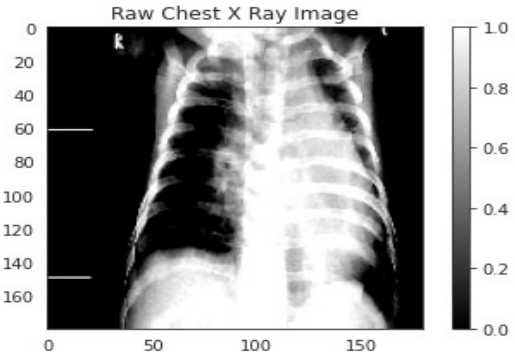


Figure 7. Raw chest X-ray image.

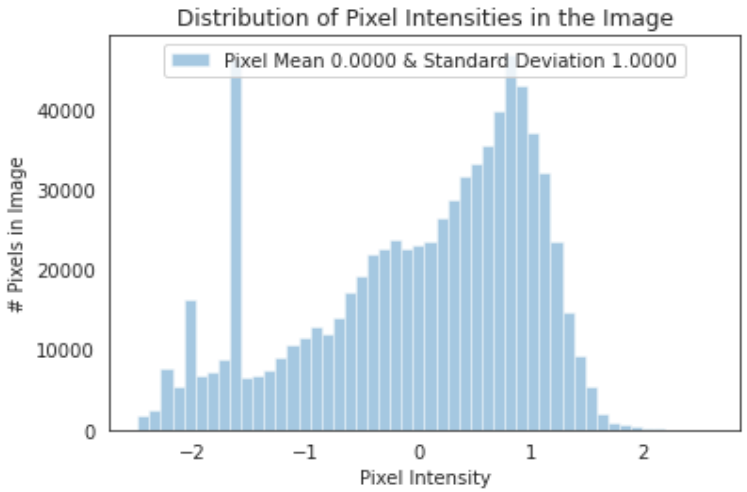


Figure 8. Pixel intensity graph.

3.3. Building a CNN Model

Table 4. Parameter types and numbers.

Parameter Types	Number of Parameters
Total Params	6,203,681
Trainable Params	6,202,785
Non-trainable Params	896

According to **Table 4**, the model has a total of 6,203,681 parameters. Of these, 6,202,785 parameters are trainable, meaning they can be changed and improved during the model training process to improve its performance in each task. On the other hand, there are 896 untrained parameters that are predicted and remain constant throughout the training period. These untrained layers usually include feature sets or pre-trained features obtained from previous examples or external sources. By separating trained and untrained phases, the model benefits from a balance between flexibility and stability. This separation allows the model to learn specific information while using pre-existing knowledge or fixed features of the structure. Controlling and understanding these parameters is important for effective model training, optimization, and obtaining optimal results in various machine learning applications. During version schooling, several parameters are frequently defined to govern and optimize the mastering procedure. The parameter "epochs" shows that the education process entails repeating the complete dataset 10 instances, allowing the model to regularly analyze from the facts. The "validation records" parameter shows that a separate validation dataset is used to check the model's performance and examine its typical capacity at some point of schooling. The "class weight" parameter suggests the mission of various weights to extraordinary instructions, allowing for example to prioritize certain instructions or to cope with an unbalanced distribution of lessons. The "steps per epoch" parameter determines the number of steps or batches to be processed in each education length, which impacts the granularity and performance of the training method. Finally, the "validation steps" parameter specifies the range of steps or agencies to be processed at some point of the validation step, controlling the frequency of the

evaluation and the computational resources used. By cautiously tuning these parameters, the version's education approach may be satisfactory tuned to obtain better overall performance, enhance connectivity, and deal with the challenges or demands of a given task.

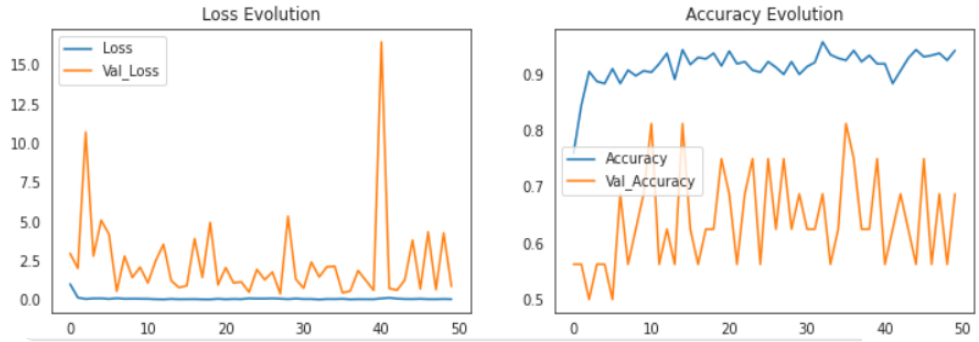


Figure 9. Loss and accuracy values of CNN Model for test and train input.

The given statistics represent the performance result of the train model after training **(Figure 9**). The first line means that the loss executed on the check dataset is 0.4553, which shows how well the model's predictions healthy the actual values, with a decrease value indicating higher overall performance. The accuracy of the corresponding look at is 83.49%, which represents the percentage of correctly anticipated labels as compared to the total range of samples inside the test dataset. The second line refers to the decrease lack of 0.1417 at the education dataset, suggesting a development in performance during training. The accuracy of the train is said to be 95.32%, which shows the share of efficaciously anticipated labels inside the training set. These results advocate that the model has carried out excessive accuracy at the train information, however there is a mild lower in performance at the check dataset, which may additionally suggest the want for similarly improvement to improve the capability to generalize the model and enhance its overall performance on unobserved information.

		Actual class	
		P	**N**
Predicted class	**P**	**91**	143
	N	21	**369**

Figure 10. Confusion matrix of CNN model.

The Confusion Matrix is a commonly used tool in machine learning to evaluate the performance and accuracy of feature classification (**Figure 10**). It provides a comprehensive overview of model predictions by comparing them with actual labels. The main purpose of using the confusion matrix is to get information about the performance of the model in different groups and to identify the sources of errors. By

analyzing these figures, we calculated various parameters such as accuracy, precision, recall, and F1 scores, which provide a clearer understanding of the model's performance.

Table 5. Evaluations obtained from the confusion matrix of CNN model.

	0	1	Accuracy	Macro Avg.	Weighted Avg.
Precision	0.95	0.81	0.84	0.88	0.86
Recall	0.62	0.98	0.84	0.8	0.84
F1-score	0.75	0.89	0.84	0.82	0.83
Support	234	390	0.84	624	624

The results of the perturbation analysis of the CNN model show that the model performed well in general (**Table 5**). The accuracy of predicting the "Normal" category (0) was excellent, at 0.95, meaning that when the model described an event as "Normal", it was 95% accurate. On the other hand, the recall for the "Normal" group was 0.62, suggesting that the model missed many "Normal" events as only 62% were correctly identified. The F1-score for the "Normal" group was 0.75, indicating a well-balanced performance in terms of precision and recall. The number of incidents in the "Normal" group was recorded as 234, which shows the support of that group. Further, the accuracy of the category "pneumonia" (1) was 0.81, suggesting that when the model predicted the event as "pneumonia", it was correct 81% of the time. The recall for the group "pneumonia" was strong, at 0.98, which means that the model correctly identified 98% of cases of "pneumonia". The "pneumonia" cluster has an F1-score of 0.89, indicating a strong combination of precision and recall. The number of cases in the "pneumonia" group was reported as 390, indicating support for that group. Accuracy measurements provide a comprehensive assessment of the model's performance. Both precision and recall were 0.84, suggesting that the model was 84% accurate. The accuracy of the F1-score was also 0.84, indicating that the entire dataset performs well in terms of accuracy and recall. Moderate support for all groups is indicated by positive support of 0.84. When statistical significance was considered, the accuracy was 0.88, suggesting acceptable overall accuracy across groups. The model has some problems in treating cases from both groups equally, as seen by the average score of 0.80. The average F1-score was 0.82, indicating a good level of accuracy and recall for all groups. The average support measured across all groups was reported as 624 with a total of major support. Finally, the weighted values provide a general assessment that accounts for group imbalance. The mean accuracy was 0.86, suggesting that the overall accuracy was satisfactory. The weighted average recall was 0.84, indicating that the model can capture samples from both groups. The weighted average F1-score was 0.83, indicating a balanced performance for all groups when precision and recall were considered. All the heavy support in all groups is indicated by the heavy support of 624. In general, the model worked well, with excellent accuracy and recall for the "pneumonia" group but significant problems collecting cases from the "Normal" group. These findings indicate that the model may require further development to accurately identify "Normal" cases while maintaining high accuracy for "pneumonia" cases.

3.3.1. DenseNet

Table 6. Parameter count of DenseNet.

Total params:	7,037,504
Trainable params:	6,953,856
Non-trainable params:	83,648

In a traditional CNN, each layer is only connected to the next layer. However, in Densenet, every layer is connected to every deep layer in the network. This dense connectivity model facilitates reuse and promotes the propagation of gradients, which can improve information flow and improve network performance. Dense connectivity in Densenet is achieved through a specific structure called a "dense block". A dense domain has many layers, where each layer takes all the previous maps as input. This design ensures that information from earlier stages is directly accessible to subsequent stages, allowing for the extraction of more efficient features. Additionally, Densenet includes a transition layer between dense blocks to control the number of feature maps. The transformation layer includes batch normalization, followed by a 1 × 1 convolutional layer and averaging. This reduces the size of feature maps, which helps reduce network complexity. In the results, the total number of parameters of the Densenet model is 7,037,504. Of these, 6,953,856 parameters are trainable, meaning they are learned during training (**Table 6**). The remaining 83,648 parameters are not configurable, which usually include batch normalization parameters or other fixed parameters. Overall, Densenet has shown remarkable performance in various computer vision tasks, such as image segmentation, object recognition, and segmentation. Its dense network has been shown to improve gradient flow, reduce the vanishing gradient problem, and promote segmentation. These characteristics make Densenet an efficient framework for deep learning tasks, leading to superior results on benchmark datasets.

```
624/624 [==============================] - 13s 21ms/step - loss: 1.3580 -
Test Accuracy: 70.99%
652/652 [==============================] - 91s 139ms/step - loss: 0.2226
Train Accuracy: 92.48%
```

Figure 11. Loss and accuracy values of DenseNet for test and train input.

In the results you provided, the Densenet model achieved an analysis loss of 1.3588 and a train loss of 0.2226 (**Figure 11**). The loss value shows how well the model works in reducing the difference between the predicted and actual values. Low loss rates usually indicate good model performance. A test accuracy of 70.99% means that the model predicted the class labels for 70.99% of the test data samples. Similarly, the training accuracy of 92.48% indicates that the model has achieved an accuracy of 92.48% on the training data. Accuracy is a measure of how well the model performs overall, with higher values indicating better performance. A confusion matrix is a useful tool for analyzing the performance of a cluster model. It shows the number of correct and incorrect guesses for each category. In the confusion matrix:

		Actual class	
		P	**N**
Predicted class	**P**	**87**	147
	N	3	**387**

Figure 12. Confusion matrix of DenseNet.

The rows represent the actual groups, while the columns represent the predicted groups. The diagonal elements of the matrix represent the correct estimates, where the predicted group matches the true group. In this case, the model correctly described 224 events of the first class and 260 of the second class. The off-diagonal elements represent misclassifications. According to the confusion matrix (**Figure 12**), it can be noted that the model did not distinguish 10 times the first group as the second group and 130 times the second group as the first group. To comment on the results, it is important to have some context about the specific task or dataset of the Densenet model that was trained and tested. However, based on the data provided, it seems that this model has achieved good accuracy in the training and test groups, although the test accuracy is slightly higher. The wrong choices shown in the confusion matrix indicate that the model makes some mistakes in distinguishing between the two groups. Further research, such as examining poorly structured samples or considering other evaluation metrics, will be necessary to understand the nature of these errors and improve the performance of the model.

Table 7. Evaluation metrics of DenseNet.

	0	**1**	**Accuracy**	**Macro Avg.**	**Weighted Avg.**
Precision	0.98	0.69	0.71	0.84	0.8
Recall	0.24	0.99	0.71	0.62	0.71
F1-score	0.39	0.81	0.71	0.6	0.66
Support	234	390	0.71	624	624

Based on the generated confusion and evaluation measures, the DenseNet model (at time 50) shows mixed performance (**Table 7)**. The accuracy of the "Normal" (0) category is as high as 0.98, suggesting that when the model predicts an image as "Normal", it is true 98% of the time. The recall for the "Normal" category, on the other hand, is very low at 0.24, which means that the model detects only 24% of "Normal" events. The accuracy of the category "pneumonia" (1) is 0.69, showing an accuracy of 69% in predicting cases of pneumonia. The recall for the "pneumonia" category is high at 0.99, which means that the model correctly identifies 99% of true pneumonia cases. The recall for the category "pneumonia" is strong, at 0.99, indicating that the model correctly detects 99% of true cases of pneumonia. The F1-score for the "Normal" group is 0.39, while it is 0.81 for the "pneumonia" group, indicating a high level of precision and recall. The overall accuracy of the model is 0.71, and the main and average metrics show good capability. According to the confusion matrix, the model correctly identified 57 cases as "Normal" in fact "Normal", but incorrectly predicted a large number of cases (177) as "Normal" is actually "pneumonia". Except

for one case (389), it was well taken care of "pneumonia". These findings suggest that the DenseNet model has a large false positive rate for "Normal" cases, indicating a potential limitation in its ability to correctly detect non-pneumonia images. More work may be needed to improve the performance of the model to correctly identify "Normal" patients while maintaining its good performance in pneumonia discrimination.

3.3.2. VGG16

Table 8. Parameter count of VGG16.

Total params:	14,714,688
Trainable params:	14,714,688
Non-trainable params:	0

```
624/624 [==============================] - 11s 18ms/step - loss: 0.3491 -
5
Test Accuracy: 84.78%
652/652 [==============================] - 92s 141ms/step - loss: 0.2014 -
13
Train Accuracy: 93.08%
```

Figure 13. Loss and accuracy values of VGG16 for test and train input.

VGG16 is a convolutional neural network (CNN) architecture that was introduced by means of the Visual Geometry Group (VGG) at the University of Oxford in 2014 (**Table 8**). It is known for its simplicity and classical design, which consists of many convolutional layers with 3 × 3 small filters, followed by max-pooling layering. The architecture also includes a dense community with hidden layers, every composed of 4096 nodes, and an output layer with 1000 nodes (Kaggle). The VGG16 architecture won reputation due to its deep layer structure, which has proven progressed overall performance in a whole lot of computer vision obligations, which include image processing, item detection, and segmentation. The use of smaller filters (3 × 3) allows a deeper community with fewer layers compared to the use of larger filters. The VGG16 model achieved a test loss of 0.3491 and a train loss of 0.2014 (**Figure 13**). Low loss rates indicate good performance, as model predictions are in good agreement with the ground truth value. It is important to note that the loss of training is lower than the loss of the test, it suggests a general level where the model works well with the training data compared to the unknown test data. The test accuracy of 84.78% indicates that the model predicted the class labels for 84.78% of the test data samples. Similarly, the training accuracy of 93.08% highlights the model's performance of 93.08% accuracy on the training data. These positive data indicate that the model's performance is good, but there is still room for improvement. To enhance the performance of the model, methods such as weight control can be considered to reduce excess weight. In addition, data augmentation techniques can be used to increase the diversity of the training data, thereby improving the overall ability of the model to generalize and predict accurately.

Table 9. Evaluation metrics of VGG16.

	0	1	Accuracy	Macro Avg.	Weighted Avg.
Precision	0.88	0.85	0.86	0.86	0.86
Recall	0.72	0.94	0.86	0.83	0.86
F1-score	0.79	0.89	0.86	0.84	0.85
Support	234	390	0.86	624	624

Precision is defined as the proportion of correctly expected cases in relation to the total number of correctly expected cases (**Table 9**). The precision of the "Normal" (0) category is 0.88, suggesting that when the model predicts an image as "Normal", it is true 88% of the time. The precision of the "pneumonia" category (1) is 0.85, showing an accuracy of 85% in predicting cases of pneumonia. The number of correctly expected cases out of the total number of positive events is measured by recall, also known as sensitivity or true positive rate. The recall for the "Normal" class is 0.72, which means that the model correctly detects 72% of "Normal" events. The recall for the "pneumonia" category is 0.94, which means that the model correctly detects 94% of all pneumonia cases. The F1-score is a balanced evaluation of model performance as it is a proportional measure of precision and recall. The F1-score for the "Normal" group is 0.79, while the F1-score for the "pneumonia" group is 0.89. An increased F1 score means better balance and recall. The number of occurrences of each group in the dataset is represented by the support. The "Normal" group has 234 supporters, while the "pneumonia" group has 390. The accuracy of the model predictions across all groups was measured. This VGG model has an accuracy of 0.86, suggesting that it predicts correctly 86% of the time. A macro average aggregates average performance across groups without considering group imbalances, but a weighted average does. The overall precision, recall, and F1 scores are respectively 0.86, 0.83, and 0.84. The weighted average precision, recall, and F1 scores are respectively 0.86, 0.86, and 0.85. A confusion matrix is a table that compares model predictions with actual labels. In this situation, the model predicted 169 cases as "Normal" actually as "Normal", and 65 cases as "pneumonia" actually as "pneumonia". It misdiagnosed 24 cases as "pneumonia" when they were actually "Normal", and correctly identified 366 cases as "pneumonia". In general, the model works very well for both groups, with high accuracy and recall, suggesting its ability to distinguish between "Normal" and "pneumonia". The effectiveness of the model is also supported by the F1 and precision scores. It is important to emphasize, however, that a comprehensive study will require more information about the specific data, the problem being addressed, and any specific methods or limitations.

3.3.3. ResNet

Table 10. Parameter count of ResNet.

Total params:	23,587,712
Trainable params:	23,534,592
Non-trainable params:	53,120

```
624/624 [==============================] - 13s 20ms/step - loss: 1.0157 - a
0
Test Accuracy: 66.67%
652/652 [==============================] - 91s 140ms/step - loss: 0.4419 -
69
Train Accuracy: 81.19%
```

Figure 14. Loss and accuracy values of ResNet for test and train input.

ResNet, short for Residual Network (**Table 10**), is a deep convolutional neural network architecture proposed by He et al. [11]. Introduced the concept of residual connections, which helped solve the problem of vanishing gradients in deep neural networks. In traditional deep systems, information flows through a series of layers, and each layer learns to extract elements from the input. However, as the network gets deeper, the gradients can become very small, making it difficult for the network to learn effectively. ResNet solves this problem by introducing skip links, also known as shortcut links or data maps.

In the results we provided for the ResNet model, the test loss of 1.0157 and the train loss of 0.4419 show how well the model works to reduce the difference between the predicted and actual values (**Figure 14**). Lower loss valuesare generally desirable, suggesting better model performance. A test accuracy of 66.67% means that the model correctly predicted the class letters for 66.67% of test data samples. Similarly, the training accuracy of 81.19% indicates that the model has achieved an accuracy of 81.19% on the training data. Higher levels of accuracy indicate better overall performance.

Comparing the accuracy of the train with the accuracy of the test, it seems that this model has a certain degree of overfitting. Overfitting occurs when a model performs well on the training data but does not fit well on the unobserved test data. Regular methods such as dropping out of school or losing weight can be used to reduce excess and improve overall.

Table 11. Evaluation metrics of ResNet.

	0	**1**	**Accuracy**	**Macro Avg.**	**Weighted Avg.**
Precision	1	0.64	0.65	0.82	0.78
Recall	0.08	1	0.65	0.54	0.65
F1-score	0.14	0.78	0.65	0.46	0.54
Support	234	390	0.65	624	624

The ResNet example (time 50) (**Table 11**) shows the performance based on the given confusion matrix and evaluation metrics. Accuracy for the “Normal” category (0) is 1, indicating that when the model predicts an image as “Normal”, it is correct 100% of the time. However, the recall for the “Normal” group is very low at 0.08, suggesting that the sample only identifies 8% of “Normal” cases. For the group “pneumonia” (1), the accuracy is 0.64, showing an accuracy of 64% in predicting cases of pneumonia. The recall for the category “pneumonia” is higher than 1, which means that the model correctly identifies all cases of pneumonia. However, the very low recall for the “Normal” category raises concerns about the model’s ability to correctly

identify “Normal” cases. The F1-score for the “Normal” group is 0.14, which shows the imbalance between precision and recall. The overall accuracy of the model is 0.65, and the main and limited metrics suggest subpar performance compared to previous models. The confusion matrix reveals that the model predicted 18 events as “Normal” in fact as “Normal”, but misclassified many events (216) as “Normal” when they were. and “pneumonia”. It correctly classified all cases (390) as “pneumonia”. These results suggest that the ResNet model struggles to accurately identify “Normal” cases and may have a high false positive rate, which may be a cause for concern in clinical settings. Further research and fine-tuning may be needed to improve the performance of the model.

3.3.4. InceptionNet

```
624/624 [==============================] - 15s 23ms/step - loss: 0.3736 - accu
racy: 0.8558 - precision: 0.8676 - recall: 0.9077
Test Accuracy: 85.58%
652/652 [==============================] - 95s 145ms/step - loss: 0.2899 - acc
uracy: 0.9043 - precision: 0.9913 - recall: 0.8790
Train Accuracy: 90.43%
```

Figure 15. Loss and accuracy values of InceptionNet for test and train input.

InceptionNet, also known as GoogLeNet, is a deep neural network architecture developed by Szegedy et al. [10]. It was designed to tackle the problems of deep network training by introducing the concept of “starting modules” and reducing the computational cost of convolutions. A key innovation in InceptionNet is the initialization module, which consists of multiple convolutional layers with different filter sizes. The purpose of the startup module is to capture information at different spatial scales by applying filters of different sizes within the same scale. This allows the network to efficiently extract local and global components.

Regarding the results we provided for the InceptionNet model, the test loss of 0.3736 and the train loss of 0.2899 show that the model has succeeded in reducing the difference between the predicted and actual values (**Figure 15**). Lower loss values are generally desirable as they indicate better performance. A test accuracy of 85.58% means that the model correctly predicted the class letters for 85.58% of the test data samples. Similarly, a train accuracy of 90.43% indicates that the model has achieved an accuracy of 90.43% on the training dataset. These accuracy scores indicate that the performance of the model is good but still leaves room for improvement.

It is vital to be aware that the outcomes we supplied show a big overall performance gap between training and test accuracy, which indicates over-performance. Overfitting occurs when a version learns to carry out properly on training data but fails to generalize to unobserved look at records. Standard methods along with dropout, weighting, or growing the size of the training dataset can be used to further reduce and improve generalization.

Table 12. Evaluation metrics of InceptionNet.

	0	1	Accuracy	Macro Avg.	Weighted Avg.
Precision	0.83	0.85	0.84	0.84	0.84
Recall	0.73	0.9	0.84	0.82	0.84
F1-score	0.77	0.88	0.84	0.82	0.84
Support	234	390	0.84	624	624

Based on the specified confusion matrix and evaluation criteria, the InceptionNet model (at 50 times) performs very well (**Table 12**). The accuracy of the "Normal" (0) category is 0.83, suggesting that when the model predicts an image as "Normal"， it is correct 83% of the time. The recall for the "Normal" class is 0.73, which means that the model correctly detects 73% of all "Normal" events. The accuracy of the "pneumonia" category (1) is 0.85, which means an accuracy of 85% in predicting cases of pneumonia. The recall for the group "pneumonia" is strong, at 0.99, which means that the model correctly detects 99% of cases of real pneumonia. Both groups have good F1 scores, with 0.77 for "Normal" and 0.88 for "pneumonia". The overall accuracy of the model is 0.84, and large and heavy metrics support this figure. According to the confusion matrix, the model predicted 170 cases as "Normal" as "Normal"， and 64 cases as "Normal" as "pneumonia". It misdiagnosed 36 cases as "pneumonia" when they were actually "Normal", and correctly identified 354 cases as "pneumonia". These findings show that the InceptionNet model performs well in classifying "pneumonia" cases but outperforms in detecting "Normal" cases compared to VGG16.

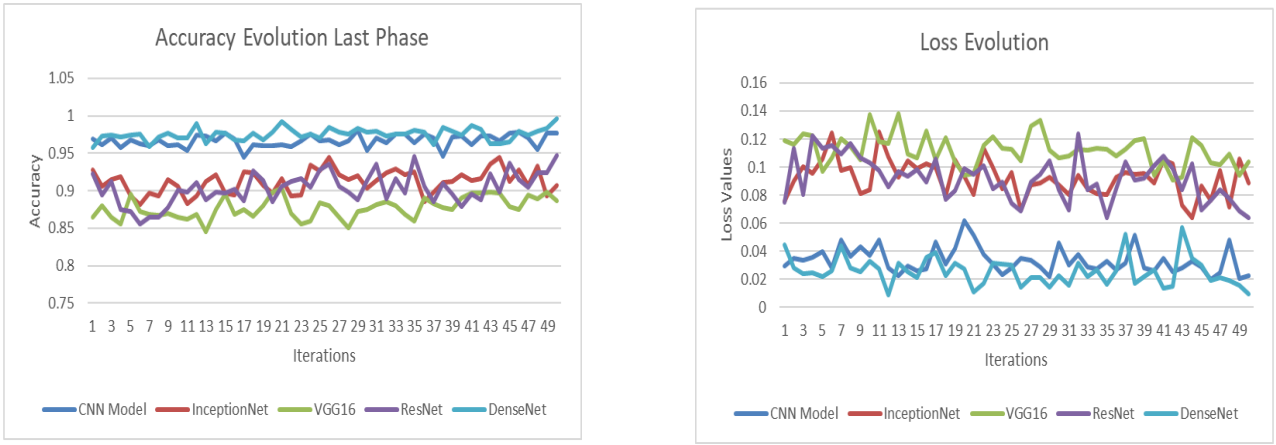


Figure 16. Final accuracy and loss evolution after 250 epochs.

DenseNet's high accuracy can be attributed to its unique dense connectivity method **(Figure 16)**. By connecting each layer to each layer in a feedforward manner, DenseNet promotes the use of layers, which can help improve gradient flow and reduce the vanishing gradient problem. This dense network allows DenseNet to accurately capture and distribute information across the network, leading to powerful performance. ResNet, with its new residual networks, has shown great accuracy in solving the vanishing edge problem. By learning residual maps instead of generating direct maps, ResNet allows deep neural networks to be trained efficiently. These

remaining connections enable the gradient to flow smoothly, simplifying the training of very deep networks and contributing to high accuracy. The VGG16, although relatively simple compared to other models, has achieved competitive accuracy. Its parallel architecture, with small convolutional filters (3 × 3) and max-pooling layering, allows capturing local features at multiple scales. However, the depth of VGG16 is relatively low compared to DenseNet and ResNet, which reduces its ability to learn complex representations and contributes to low accuracy. InceptionNet, particularly Inception V3, has slightly lower accuracy compared to the others. The Inception architecture, with its inception modules, captures information at different spatial scales using parallel convolutional filters. While this design enables computational efficiency, it introduces increased complexity, leading to challenges in training and potentially impacting accuracy.

4. Conclusion

Through multiple training phases with the same input, the models were able to achieve increased accuracy. This iterative process allowed the models to learn and refine their predictions, resulting in improved performance over time. Despite the limitation of limited computing power and server space, the models underwent a maximum of 250 training iterations (epochs). This constraint was necessary to balance the computational resources available while still achieving notable progress. Additionally, it is worth noting that reviews from the literature support the notion that these models are robust and generalizable. Therefore, even when tested with different X-ray image inputs, the models are expected to exhibit strong performance, indicating their reliability and effectiveness across various scenarios.

The difference in accuracy among the models is attributed to the fundamental design variations between them.

Author contributions: Conceptualization, AN and RSD; methodology, SBS; writing—original draft preparation, EU and SIS; writing—review and editing, IK, SIS and EU; supervision, RSD and SST. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Baque-Juston M, Pellegrin A, Leroy S, et al. Organizing pneumonia: What is it? A conceptual approach and pictorial review. In Diagnostic and Interventional Imaging. Elsevier Masson SAS. 95(9): 771–777. doi: 10.1016/j.diii.2014.01.004
2. Ayan E, Ünver HM. Diagnosis of Pneumonia from Chest X-Ray Images using Deep Learning. In: Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT); 24−26 April 2019; Istanbul, Turkey.
3. Jaiswal AK, Tiwari P, Kumar S, et al. Identifying pneumonia in chest X-rays: A deep learning approach. Measurement: Journal of the International Measurement Confederation. 2019; 145: 511–518. doi: 10.1016/j.measurement.2019.05.076
4. Szepesi P, Szilágyi L. Detection of pneumonia using convolutional neural networks and deep learning. Biocybernetics and Biomedical Engineering. 2022; 42(3): 1012–1022. doi: 10.1016/j.bbe.2022.08.001
5. Elshennawy NM, Ibrahim DM. Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images. Diagnostics. 2020; 10(9). doi: 10.3390/diagnostics10090649

6. Mohammed Ahmed A, Alhadi Babikir G, Mohammed Osman S. Classification of Pneumonia Using Deep Convolutional Neural Network. American Journal of Computer Science and Technology. 2022; 5(2): 26. doi: 10.11648/j.ajcst.20220502.11
7. Varshni D, Nijhawan R, Thakral K, et al. Pneumonia Detection Using CNN based Feature Extraction. In: Proceedings of the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT); 20−22 February 2019; Coimbatore, India.
8. Hassan M. ul. VGG16—Convolutional Network for Classification and Detection. Available online: https://neurohive.io/en/popular-networks/vgg16/ (accessed on 9 May 2024)
9. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Computer Science. 2017.
10. Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions. Computer Science. 2014.
11. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Computer Science. 2015.
12. Desai P, Pujari J, Sujatha C, et al. Hybrid Approach for Content-Based Image Retrieval using VGG16 Layered Architecture and SVM: An Application of Deep Learning. SN Computer Science. 2021; 2(3). doi: 10.1007/s42979-021-00529-4
13. Tao Y, Xu M, Lu Z, Zhong Y. Dense net-based depth-width double reinforced deep learning neural network for high-resolution remote sensing image per-pixel classification. Remote Sensing. 2018; 10(5). doi: 10.3390/rs10050779
14. Tiwari A. Supervised learning: From theory to applications. Artificial Intelligence and Machine Learning for EDGE Computing. 2022; 23–32. doi: 10.1016/B978-0-12-824054-0.00026-5

Imaging and Radiation Research 2024, 7(1), 6404.
https://doi.org/10.24294/irr6404

Perspective

Offshore reporting of radiologic examinations supplementing healthcare delivery worthy of Medicare reimbursement

Arjun Kalyanpur[1], Neetika Mathur[2,*]

[1] Teleradiology Solutions, Bengaluru, Karnataka 560048, India
[2] Image Core Lab, Whitefield, Bengaluru, Karnataka 560048, India
*** Corresponding author:** Neetika Mathur, neetika.mathur@imagecorelab.com

CITATION

Kalyanpur A, Mathur N. (2024). Offshore reporting of radiologic examinations supplementing healthcare delivery worthy of Medicare reimbursement. Imaging and Radiation Research. 7(1): 6404. https://doi.org/10.24294/irr6404

ARTICLE INFO

Received: 15 May 2024
Accepted: 30 May 2024
Available online: 19 June 2024

Abstract: Medicare, a major healthcare program under the Centers for Medicare & Medicaid Services (CMS) has extended telemedicine services within several states in the US for different specialties for which it reimburses in order to establish a qualitative and accessible healthcare system. In parallel, it has been seen that teleradiology services by American Board Certified radiologists based offshore can significantly supplement healthcare delivery in the US by mitigating the shortage of radiologists and enhance outcomes of patient care especially for after-hours emergency work. Teleradiology can help workflow by improving workload distribution, lowering the cost of reporting, shortening turn-around-time for reports, and improving quality of life for staff. The aim of the article is to provide perspective on Medicare reimbursement of offshore telereporting services. We submit that due to its value proposition and contribution to healthcare, offshore telereporting by American Board Certified Radiologists is worthy of Medicare reimbursement and should be re-evaluated for its credits.

Keywords: Medicare; teleradiology; offshore; reimbursement; healthcare

1. Introduction

Radiology is a medical specialty that has become a primary contributor to human healthcare [1]. It involves the acquisition and interpretation of images of the human body for the diagnosis of a number of diseases and abnormalities. Technological innovation paved the way for teleradiology, which involves the electronic transmission of diagnostic imaging studies such as X-rays, CTs, and MRIs to remote sites for consultation or interpretation. Teleradiology, a subset of telemedicine, has played a significant role in delivering high quality contemporaneous radiologic interpretations, particularly in areas or during time periods where there is a shortage of radiologists, to facilitate emergency consultations and improve standards of patient care. It has been considered as a front-line driver in making digital imaging achieve its deserved potential.

The benefits of teleradiology for patients have been well documented in several studies [1–5]. In terms of scale of use, in the United States in 2014, more than 50% of all telemedicine services were reported to be performed by teleradiology [6]. The data from the American Medical Association's 2016 Physician Practice Benchmark Survey reveals that physicians practicing in radiology (39.5%), psychiatry (27.8%), and cardiology (24.1%) frequently use telemedicine to connect with patients. Radiologists (25.5 percent) are in the third position among all specialties, to use telemedicine to connect with other health care professionals (having a specialty consultation and

getting a second opinion) after emergency medicine physicians (38.8 percent) and pathologists (30.4 percent) [7].

For teleradiology utilization, either the images are transmitted from the hospital to the residence of the hospital-based on-call radiologist in the United States after usual working hours or transmitted to a commercial teleradiology service provider that employs American Board Certified radiologists located in other states within US or offshore to carry out preliminary or final interpretations. However, in the latter case, the American Board Certified radiologists located offshore i.e., in countries other than the United States have been permitted to render only preliminary readings and not final radiologic interpretations. In these cases, onsite radiologists overread the images the next day and provide a final interpretation. This model has been previously validated and described in the literature [2,8,9].

The Centers for Medicare & Medicaid Services (CMS), an agency within the US Department of Health and Human Services (HHS) governs the nation's major healthcare programs including Medicare, Medicaid, the Children's Health Insurance Program (CHIP), and The State and Federal health insurance marketplaces. It evaluates the amassed data and prepares research reports, operates to remove the cases of fraud within the healthcare system, and manages the payments for all radiology services [10]. It decides reimbursement rates for all medical services and equipment covered under Medicare. The services are required to be medically essential, be ordered by physicians, and have documentation to support the submitted claims [11]. Generally, Medicare is available for people age 65 or older and people with disabilities and chronic conditions. Medicare has two parts, Part A (hospital insurance) and Part B (Medicare insurance). Medicare Part B helps cover medical services such as doctors' services, outpatient care, and other medical services including teleradiology services (discussed in Pub. 100-02, Medicare Benefit Policy Manual, chapter 15, section 30). The interpretation of an X-ray, electrocardiogram, electroencephalogram, etc. are enlisted examples [12]. The cost of radiology comprises the technical fees related to the acquisition of images including the fee for operating the devices and paying the radiology technologists as well as the radiologist's fees for reading and interpreting the images. Charges differ depending on the type of modality (e.g., MRI, CT), on whether contrast is used or not, on the body part/organ (e.g., breast, head, leg), and whether there is an interventional procedure or not [1].

1.1. Offshore reporting of radiologic examinations supplementing healthcare delivery

Teleradiology services located within the United States have been working proficiently but face difficulty recruiting radiologists for night-time working hours [13]. Additionally, from an economic perspective the radiologists working nights are inherently unproductive and represents a significant cost burden to the healthcare system given that the current standard/expectation is typically 'one week on one week off' or often 'one week on and two weeks off' to allow for physician recovery from the unphysiological lifestyle and sequelae of night shift work. Furthermore, nightshift work is, for obvious reasons, perceived as being unattractive, rendering recruitment to this cohort especially challenging. Offshore teleradiology has demonstrated the

potential to address this problem and deliver quality and timely radiological interpretations through night-shift teleradiology services delivered by US Board certified radiologists when onsite radiologists are unable to provide immediate coverage [14,15]. Various 'nighthawk' teleradiology groups have evolved by leveraging the growing opportunities that teleradiology presents [16,17].

A survey was conducted to determine the effects of international teleradiology attending radiologist coverage (ITARC) of emergency examinations on radiology residents' perceptions of night call. ITARC is the time gap when a teleradiologist is awake and work for normal daytime hours, at the same time covering the night shift in the US. Most surprisingly, the survey results revealed that ITARC relieved radiology residents' stress and anxiety related to on-call shifts and promoted accurate afterhours readings and availability of attending radiologists for consultation with referring clinicians, reduced load on daytime attending radiologists and enhanced their educational experience as well. However, ITARC necessitates licensure and credentialing of off-shore teleradiologist in US hospitals, a secure network, redundant internet connections to banish downtime and an expeditious transmission of images for contemporaneous interpretations [18].

The benefits of 'nighthawk' services were also revealed by Goelman [19]. The study reported that 'nighthawk' services rendered through teleradiology supported by quick and secure internet connections resulted in enhanced night-time radiologist productivity, better quality of life, and most significantly, high quality radiology interpretations.

Furthermore, burnout, a global health problem, is also prevalent among US physicians including radiologists. Numerous studies have reported that burnout is a cluster of symptoms developing from severe work-related stress, apparent as emotional fatigue, depersonalization, despondence, and lethargy [20,21]. It can also lead to reduction in physician productivity, professional effort, gratification, impaired performance and may even result in elevated physician turnover, early retirement contributing to worsening physician shortages, and tragically even physician suicide, thus eventually leading to increasing health care costs. A study by Canon et al. [22] revealed that 54%–72% of diagnostic radiologists and interventional radiologists under study reported aforesaid symptoms of burnout. Thus, the utilization of off-shore teleradiology services addresses burnout, improves workload distribution, reduces the diagnostic error rate, shortens turn-around-time for reports, and enhances the quality of life for radiologists. This has been well documented in various published studies of teleradiology [2,16,23–30].

Unfortunately, despite its manifested value proposition, offshore telereporting has still not received the desired credit for its contribution to healthcare. In the United States, Medicare and Medicaid laws prohibit radiologists who are located in countries other than the United States to qualify for reimbursement for final reads. Broadly, Medicare will not pay for health care or supplies that are conducted outside the United States (US). The term "outside the US" means anywhere other than the 50 states of the US, the District of Columbia, Puerto Rico, the US Virgin Islands, Guam, American Samoa, and the Northern Mariana Islands (discussed in Pub. 100-02, chapter 16, section 60, for exceptions to the "outside the US" exclusions) [12]. For this reason, offshore radiology reports are delivered in the preliminary or wet-read model which

necessitates subsequent review by an onshore radiologist (typically at the hospital of origin of the images). This results in duplication of effort and further strains a system that is already overwhelmed and subject to challenges such as reporting delays, reporting errors and radiologist burnout.

Interestingly, the ACR Task Force on International Teleradiology, in 2005, released a white paper with the aim of addressing the legal, regulatory, reimbursement, insurance, quality assurance, and other issues related to international teleradiology. The task force acknowledged that there is no technological variation between intrinsically or offshore generated teleradiology interpretations and reports. In either case, quality and competency should be the priority. Worthy of mentioning, the task force also strongly opined that ABR certified status is the most trustworthy parameter for the quality of an interpreting physician. Moreover, reimbursement for radiologic interpretations and ensuing reports that are furnished by international teleradiology is predicated upon the expectation that the radiologists must be certified by the American Board of Radiology, should have medical licenses in every state and hold privileges, credentialed as a member of the medical staff and have professional liability insurance coverage at the institution or hospital performing the examination and receiving the report [13,31].

The confirmation of professional standing by way of medical licensure and credentialing of radiologists empaneled by teleradiology service providers, as well as stringent quality assurance programs, are pivotal in designing the outcomes of the use of teleradiology to offshore radiology services [32]. Moreover, the advent and integration of PACS (picture archiving and communication system) and RIS (radiology information system) into the teleradiology system, ensued proficient transmission of imaging and findings between teleradiologist and referring clinician [33]. An article reported that hundreds of US hospitals utilize overseas or offshore teleradiology services rendered by the teleradiology service providers such as teleradiology solutions, Bangalore, which strictly follow ACR guidelines regarding licensure, insurance, and hospital privileges. However, Medicare laws prohibit reimbursement to such offshore providers [16]. Besides reading images per se, some international teleradiology firms are also performing 3D image reconstruction for US hospitals [33].

The American College of Radiology [34], together with the American Association of Physicists in Medicine and the Society for Imaging Informatics in Medicine, issued an upgraded ACR technical standard for the electronic practice of medical imaging which clearly described the objectives and adequacy for the utilization of digital image data, along with the electronic transmission of patient examinations from one location to another for interpretation. In 2013, a White Paper of ACR Task Force on International Teleradiology recognized the role of teleradiology in patient care, in ameliorating access to radiologic services and subspecialty expertise in areas in which it is otherwise unavailable. The white paper also recognized the need for designing protocols and software for better connections between physicians, technologists and patients, rules for sharing electronic medical record and peer review system and thus refined the guidelines and standards for teleradiology practice focusing on the specified concerns [35].

In 2019, a survey was carried out by the ACR Commission on Human Resources Workforce to determine the constitution of the radiology workforce and understand the current job market for radiologists. The results indicated that 8% of the workforce is greater than 65 years of age and 22% are between 56 and 65 years [36]. In another study, among 20,970 radiologists involved in active patient care, 82% were of age 45 and over, while 53% were age 55 and over [37]. This indicated that the future workforce needs will depend on retirements of these senior radiologists. In a study presented at RSNA 2021, Khurana et al reported that the increase in the Medicare population surpassed the diagnostic radiology (DR) workforce by about 5% from 2012 to 2019. Further, the pipeline of the incoming radiologist is not commensurate with the need, as from 2010 to 2020, the number of DR trainees entering the workforce increased by 2.5% as compared to a 34% rise in the number of adults over 65. The study by Khurana et al also projected a 4.2 times rise in the number of radiologists per 100,000 Medicare enrollees from 2012 to 2019 in US [38,39]. A salve for these current workforce problems is teleradiology services provided by off-shore radiologists which can add to the capacity of American Board Certified radiologists and help bridge the shortfall.

1.2. Medicare reimbursement

Medicare has implemented strict guidelines through which it will reimburse telemedicine practices. The eligibility for Medicare reimbursement for a telemedicine service depends upon the patient's location. The patient must be in a rural geographical location either a health professional shortage area or a county outside of a metropolitan statistical area with exceptions for patients getting treatment for end-stage renal disease, stroke, and substance use disorder [40]. Medicare makes payments under the physician fee schedule (PFS) for the services of more than 10,000 physician services and other billing professionals (i.e., payment of assistant at surgery, nurse practitioners, nurse midwives, physician assistant, clinical psychologists and social workers, registered dietitians or nutrition professionals etc.), since 1992. The Medicare PFS pricing amounts are adjusted to display the difference in practice costs from area to area. Under the PFS, the payment for the physicians' services is conferred under a variety of settings, including physician offices, hospitals, critical access hospitals, skilled nursing facilities and other post-acute care settings, outpatient dialysis facilities, clinical laboratories, and beneficiaries' homes [41].

A national private payer reimbursement online survey conducted by the American Telemedicine Association interpreted that there was no standard protocol for billing for telemedicine services in the hospitals because neither government nor private payers were willing to pay for them. Moreover, insurance companies followed the guidelines of their individual states. Administrative rules varied for in-person and telemedicine care which put impediments to reimbursement. It was postulated that the setting up of universal coverage policies by regulatory bodies would remove these barriers [42]. This approach is likely to be of greater benefit given that the challenges of radiologist shortages are neither local nor regional but rather national. The increasing utilization of telemedicine has resulted in raising interest among various

payers, be it insurance companies, or certain government-funded programs, to expand their policies to accommodate for teleservices.

In 2018, Medicaid widened the scope of telehealth and telemedicine services in several states within the US for which they reimburse, thus lowering impediments to their use. Despite support from lawmakers, administrators, and clinicians in favour of continued utilization of telehealth after the COVID-19 pandemic, there is ongoing debate as to whether telehealth will continue to be reimbursed in parity with in-person care [43]. There is however no dearth of legislation related to potentially improving healthcare reimbursement practices. For example. CMS had decided on its regulations to show required changes in telehealth reimbursements made by the Bipartisan Budget Act of 2018, specifically related to end-stage renal disease (ESRD) services and the treatment of acute stroke, with effect from January 2019 [44]. According to a report by American Society of Radiologic Technologists, on 1 June 2021, the Medicare Access to Radiology Care Act [45] was introduced by US Reps. Mike Doyle of Pennsylvania and John Curtis of Utah as House Resolution 3657 with companion legislation, Senate Bill 2641, introduced on 5 August by Sen. John Boozman and cosponsors Sen. Bob Casey of Pennsylvania and Sen. Steven Daines of Montana. These bills propose a law that revises Medicare reimbursement policy for radiologist assistants to bring it at par with state radiologist assistant licensure laws essentially recognizing that innovative approaches are needed to address these critical radiologist shortages. Additionally certain coverage restrictions around PET imaging outside of oncology were lifted by CMS in July 2021 [46]. However, the 2024 MPFS puts forth new difficulties for radiologists through reimbursement reductions and the pause of the appropriate-use criteria (AUC program) for advanced diagnostic imaging services initiated in 2014 [47–49].

In summary, a number of ground-breaking legislations have been passed in recent days to support telemedicine reimbursement that will positively impact on healthcare budgets and spending. However, off-shore teleradiology still awaits its legitimate credit for the value it provides.

2. Conclusion

Medicare has expanded the reach of telehealth and telemedicine services in several states within the US for different specialties for which they reimburse, to create a qualitatively superior healthcare system that is more accessible, affordable, and empowered. Our submission is that despite this, and despite the multiple obvious stated benefits of the offshore model, offshore teleradiology delivered by American Board Certified Radiologists still does not receive its due credit. We would submit that the night-to-day international teleradiology model, two decades on from its inception, represents a successful model that deserves commensurate attention from the standpoint of reimbursement. Essentially this is an idea whose time has come.

3. Take home points

1) The Medicare regulation restricting reimbursement for healthcare services delivered overseas dates back to a time when it was intended to deter individuals from travelling overseas for procedures performed by international physicians and then submitting claims for reimbursement. The regulation did not take into account telemedicine services, which were not available at the time.
2) Today, given severe radiologist shortages in the US, and resultant radiologist overwork and burnout, American Board Certified radiologists located offshore can significantly support and supplement the healthcare delivered by the local radiologists in the US, especially for after-hours work, which can be more physiologically performed in a daylight time zone.
3) The virtual pool of radiologists available through teleradiology increases the doctor-patient ratio compensating for radiologist shortage especially at the time of emergency situations.
4) Given these benefits, and given than Medicare has been making innovative changes within the billing framework overall, it seems relevant that the issue of Medicare reimbursement for radiology reporting services delivered from offshore by American Board Certified Radiologists should be re-evaluated at this time, as this has the potential to address the challenges of shortages of radiologists which are being acutely perceived at this time.

Author contributions: Conceptualization, AK and NM; writing—original draft preparation, NM; writing—review and editing, AK; visualization,AK; supervision, AK; project administration, AK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Bashshur RL, Krupinski EA, Thrall JH, Bashshur N. The Empirical Foundations of Teleradiology and Related Applications: A Review of the Evidence. Telemedicine and E-Health. 2016; 22(11): 868-898. doi: 10.1089/tmj.2016.0149
2. Kalyanpur A, Neklesa VP, Pham DT, et al. Implementation of an international teleradiology staffing model. Radiology. 2004; 232(2): 415-419. doi: 10.1148/radiol.2322021555
3. Kalyanpur A, Meka S, Joshi K, et al. Teleradiology in Tripura: Effectiveness of a Telehealth Model for the Rural Health Sector. International Journal of Health Technology and Innovation. 2022; 1: 7-12.
4. Bogner P, Chadaide Z, Lenzsér G, et al. Teleradiology-based stroke network in Western and Southern Transdanubia in Hungary. Orvosi Hetilap. 2021; 162(17): 668-675. doi: 10.1556/650.2021.32097
5. Kiuru MJ, Paakkala TA, Kallio TT, et al. Effect of teleradiology on the diagnosis, treatment and prognosis of patients in a primary care centre. Journal of Telemedicine and Telecare. 2002; 8(1): 25-31. doi: 10.1258/1357633021937424
6. Weinstein RS, Lopez AM, Joseph BA, et al. Telemedicine, Telehealth, and Mobile Health Applications That Work: Opportunities and Barriers. The American Journal of Medicine. 2014; 127(3): 183-187. doi: 10.1016/j.amjmed.2013.09.032
7. Kane CK, Gillis K. The Use of Telemedicine by Physicians: Still the Exception Rather Than the Rule. Health Affairs. 2018; 37(12): 1923-1930. doi: 10.1377/hlthaff.2018.05077
8. Kalyanpur A, Weinberg J, Neklesa V, et al. Emergency radiology coverage: Technical and clinical feasibility of an international teleradiology model. Emergency Radiology. 2003; 10(3): 115-118. doi: 10.1007/s10140-003-0284-5

9. Abujudeh HH. Malpractice and Radiology: A Hapless Relationship. Radiology Noninterpretive Skills: The Requisites. 2018; 27: 256-266.
10. Kagan J. Centers for Medicare & Medicaid Services (CMS): Definition, How It Works. Available online: https://www.investopedia.com/terms/u/us-centers-medicare-and-medicaid-services-cms.asp (accessed on 3 March 2024).
11. Ghoshal M. Understanding Medicare Reimbursement & Claims. Available online: https://www.healthline.com/health/medicare/medicare-reimbursement#:~:text=Medicare%20pays%20for%2080%20percent,some%20of%20the%2020%20percent (accessed on 3 March 2024).
12. Rachele B. Navigating Medicare Coverage Outside the US and Foreign Travel. Available online: https://policyscout.com/medicare/learn/medicare-outside-the-us (accessed on 3 March 2024).
13. Bradley WG. Offshore Teleradiology. Journal of the American College of Radiology. 2004; 1(4): 244-248. doi: 10.1016/j.jacr.2003.12.043
14. Kalyanpur A. Commentary: Teleradiology: The Indian Perspective. Indian Journal of Radiology and Imaging. 2009; 19(01): 19-20. doi: 10.4103/0971-3026.45338
15. Kalyanpur A. The Role of Teleradiology in Emergency Radiology Provision. Health management. 2014; 14(1).
16. Wachter RM. International Teleradiology. New England Journal of Medicine. 2006; 354(7): 662-663. doi: 10.1056/NEJMp058286
17. Burute N, Jankharia B. Teleradiology: The Indian perspective. Indian Journal of Radiology and Imaging. 2009; 19(01): 16-18. doi: 10.4103/0971-3026.45337
18. Joffe SA, Burak JS, Rackson M, et al. The Effect of International Teleradiology Attending Radiologist Coverage on Radiology Residents' Perceptions of Night Call. Journal of the American College of Radiology. 2006; 3(11): 872-878. doi: 10.1016/j.jacr.2006.02.014
19. Goelman A. Telework That Works: Teleradiology and the Emergence of Nighthawk Radiology Firms. In: Proceedings of the Sloan Foundation Industry Studies Annual Conference; 2007; Cambridge, Massachusetts.
20. Shanafelt TD, Hasan O, Dyrbye LN, et al. Changes in Burnout and Satisfaction with Work-Life Balance in Physicians and the General US Working Population Between 2011 and 2014. Mayo Clinic Proceedings. 2015; 90(12): 1600-1613. doi: 10.1016/j.mayocp.2015.08.023
21. Chetlen AL, Chan TL, Ballard DH, et al. Addressing Burnout in Radiologists. Academic Radiology. 2019; 26(4): 526-533. doi: 10.1016/j.acra.2018.07.001
22. Canon CL, Chick JFB, DeQuesada I, et al. Physician Burnout in Radiology: Perspectives from the Field. American Journal of Roentgenology. 2022; 218(2): 370-374. doi: 10.2214/AJR.21.26756
23. Zafar SR. Teleradiology 2.0: How AI is changing the game in radiology platforms. Available online: https://www.linkedin.com/pulse/teleradiology-20-how-ai-changing-game-radiology-platforms-zafar/ (accessed on 3 March 2024).
24. Horn B, Chang D, Bendelstein J, Hiatt JC. Implementation of a teleradiology system to improve after-hours radiology services in kaiser Permanente southern California. The Permanente Journal. 2006; 10(1): 47-50. doi: 10.7812/TPP/05-119
25. Matsumoto M, Koike S, Kashima S, Awai K. Geographic Distribution of Radiologists and Utilization of Teleradiology in Japan: A Longitudinal Analysis Based on National Census Data. PLOS ONE. 2015; 10(9): e0139723. doi: 10.1371/journal.pone.0139723
26. Lester N, Durazzo T, Kaye A, et al. Referring Physicians' Attitudes Toward International Interpretation of Teleradiology Images. American Journal of Roentgenology. 2007; 188(1): W1-W8. doi: 10.2214/AJR.05.1303
27. Zabel AOJ, Leschka S, Wildermuth S, et al. Subspecialized radiological reporting reduces radiology report turnaround time. Insights into Imaging. 2020; 11(1): 114. doi: 10.1186/s13244-020-00917-z
28. Kaye AH, Forman HP, Kapoor R, Sunshine JH. A Survey of Radiology Practices' Use of After-Hours Radiology Services. Journal of the American College of Radiology. 2008; 5(6): 748-758. doi: 10.1016/j.jacr.2008.01.009
29. DeCorato DR, Kagetsu NJ, Ablow RC. Off-hours interpretation of radiologic images of patients admitted to the emergency department: Efficacy of teleradiology. AJR. American Journal of Roentgenology. 1995; 165(5): 1293-1296. doi: 10.2214/ajr.165.5.7572522

30. Wong WS, Roubal I, Jackson DB, et al. Outsourced teleradiology imaging services: An analysis of discordant interpretation in 124,870 cases. Journal of the American College of Radiology: JACR. 2005; 2(6): 478-484. doi: 10.1016/j.jacr.2004.10.013
31. Van Moore A, Allen B, Campbell SC, et al. Report of the ACR task force on international teleradiology. Journal of the American College of Radiology. 2005; 2(2): 121-125. doi: 10.1016/j.jacr.2004.08.003
32. Levy FS. The International Teleradiology Industry: Successes and Failures. In: Proceedings of the IHEA 2007 6th World Congress: Explorations in Health Economics Paper; 2007.
33. Levy F, Yu KH. Offshoring Radiology Services to India. Available online: http://isapapers.pitt.edu/148/1/2007-33_Yu.pdf (accessed on 3 March 2024).
34. American College of Radiology. ACR-AAPM-SIIM Technical Standard for Electronic Practice of Medical Imaging. Available online: https://www.acr-/media/ACR/Files/Practice-Parameters/Elec-Practice-MedImag.pdf (accessed on 3 March 2024).
35. Silva E, Breslau J, Barr RM, et al. ACR White Paper on Teleradiology Practice: A Report from the Task Force on Teleradiology Practice. Journal of the American College of Radiology. 2013; 10(8): 575-585. doi: 10.1016/j.jacr.2013.03.018
36. Bender CE, Bansal S, Wolfman D, Parikh JR. 2019 ACR Commission on Human Resources Workforce Survey. Journal of the American College of Radiology. 2020; 17(5): 673-675. doi: 10.1016/j.jacr.2020.01.012
37. Montecalvo R. Radiologist Staffing Trends. Available online: https://blog.vrad.com/radiologist-staffing-trends-2021 (accessed on 3 March 2024).
38. Khurana A, Patel B, Sharpe R. Geographic Variations in Growth of Radiologists and Medicare Enrollees From 2012 to 2019. Journal of the American College of Radiology. 2022; 19(9): 1006-1014. doi: 10.1016/j.jacr.2022.06.009
39. Henderson M. Radiology Facing a Global Shortage. Available online: https://www.rsnanews/2022/may/Global-Radiologist-Shortage (accessed on 3 March 2024).
40. Hyder MA, Razzak J. Telemedicine in the United States: An Introduction for Students and Residents. Journal of Medical Internet Research. 2020; 22(11), e20839. doi: 10.2196/20839
41. Medical Billers and Coders. Available online: https://www.medicalbillersandcoders.com/articles/best-billing-and-coding-practices/Understanding-Basics-of-Physician-Fee-Schedule-PFS.html (accessed on 24 March 2024).
42. Antoniotti NM, Drude KP, Rowe N. Private Payer Telehealth Reimbursement in the United States. Telemedicine and E-Health. 2014; 20(6): 539-543. doi: 10.1089/tmj.2013.0256
43. Ellimoottil C. Understanding the Case for Telehealth Payment Parity. Available online: http://www.healthaffairsdo/10.1377/forefront.20210503.625394/full/ (accessed on 3 March 2024).
44. Telehealth Policy 101—CCHP. Available online: https://www.cchpcapolicy-101/ (accessed on 3 March 2024).
45. The Medicare Access to Radiology Care Act (MARCA). American Society of Radiologic Technologists. Available online: https://www.asrtmain/standards-and-regulations/legislation-regulations-and-advocacy/marca (accessed on 3 March 2024).
46. Stempniak M. CMS dropping coverage restrictions around PET imaging outside of oncology care. Available online: https://radiologybusiness.com/topics/healthcare-policy/cms-coverage-restrictions-pet-imaging-oncology (accessed on 24 March 2024).
47. 2024 MPFS Final Rule: Impact on Radiologists. Available online: https://apsmedbill.com/whitepapers/2024-mpfs-final-rule-impact-radiologists (accessed on 4 March 2024).
48. 2024 Medicare Physician Fee Schedule Final Rule Includes Payment Reductions for Radiology. Available online: https://info.hapusa.com/blog-0/medicare-fee-schedule-2024-final-rule-includes-payment-reductions-for-radiology (accessed on 4 March 2024).
49. Schartz E, Manganaro M, Schartz D. Declining Medicare Reimbursement for Diagnostic Radiology: A 10-Year Analysis Across 50 Imaging Studies. Curr Probl Diagn Radiol. 2022; 51(5): 693-698. doi: 10.1067/j.cpradiol.2022.01.007

EnPress Publisher, LLC

Add: 9650 Telstar Avenue, Unit A, Suite 121, El Monte, CA 91731, USA.
Email: contact@enpress-publisher.com
Web: https://systems.enpress-publisher.com

ISSN 2578-1618 01
9 772578 161070