A Lightweight Residual Network For Pneumonia Classification

Yao Lu
ShiHeZi University
XinJiang, China
18709931890@163.com

YuChen Zheng*
ShiHeZi University
XinJiang, China
ouczyc@outlook.com

Abstract

Pneumonia, a global health disease, continues to threaten human life. In this challenge, accurate and rapid diagnosis not only contributes to treatment effectiveness, but also reduces the burden on the healthcare system. With the development of medical imaging technology, chest X-ray image has gradually become an important tool in the diagnosis of pneumonia. However, the surge in the number of pneumonia patients has challenged traditional diagnostic methods with time and resource constraints. Deep learning, particularly convolutional neural networks (CNNs), shows potential for problem solving in this context. However, the traditional CNN model has large parameters and is not suitable for resource-constrained devices. Therefore, this study proposes a lightweight residual network (LR-Net) that combines residual network (ResNet) and depth-separable convolution (DWConv) for early and automated detection of pneumonia in chest X-ray images. Focusing on the binary classification problem (pneumonia and normal), the model was tested on the ChestX-ray2017 dataset with an accuracy of 86.49%, model parameter of 268.098 K and FLOPs of 732.2184 M. This lightweight model greatly reduces the demand for computing resources and memory. Making it ideal for mobile devices and edge computing devices. Taken together, the findings of this study demonstrate the effectiveness of lightweight deep learning models combining ResNet and DWConv in pneumonia diagnosis, providing clinicians with a powerful tool to enhance diagnostic insight into pneumonia. In addition, due to its lightweight nature, the model offers new possibilities for the rapid and accurate diagnosis of pneumonia in resource-constrained settings, helping to reduce the strain on global healthcare systems and provide patients with timely treatment.

Keywords—Pneumonia classification, Deep learning, LR-Net

I. INTRODUCTION

Pneumonia continues to pose a major threat to global health.Rapid and accurate diagnosis of pneumonia is not only essential for effective treatment, but also plays a crucial role in reducing the burden on the healthcare system [1]. Advances in medical imaging technology have made chest X-ray images an important diagnostic tool for detecting pneumonia. However, the surge in pneumonia cases in the post-pandemic era poses a serious challenge to traditional diagnostic methods, which are time-consuming and constrained by limited resources[2].

In this context, deep learning, particularly convolutional neural networks (CNNs), has emerged as a powerful solution. They show great potential for solving complex problems with high precision and efficiency [3][4][5][6]. At present, many studies have used CNN to classify pneumonia. Wu et al.[7] proposed a hybrid deep learning model for multi-label pneumonia image classification. Thakur et al. [8] based on traditional CNN, covid-19 is classified into two categories. Xu et al.[9] put forward the similarity regularization from contrast learning to enable CNN to learn more efficient representation of parameters, thus improving the accuracy and sensitivity of CNN. Tan et al.[10] reconstructed images based on super resolution and VGG16 neural network are used for classification of chest CT images. However, the CNN parameters in the above study are of a large scale, so they are not suitable for deployment on devices with limited computing resources[7]. In addition, with the deepening of network depth, feature degradation may occur.

In order to address these challenges, this study proposes a lightweight residual network (LR-Net), the residual network [12] and depthwise convolution [13] synergy appropriate precautions，i.e. The residual block in ResNet is replaced with a 3×3 DWConv layer and a residual connection, which realizes the model lightweight while the features do not degenerate as the depth of the network deepens. LR-Net is designed for the early automatic detection of pneumonia from chest X-ray images and is primarily used for binary classification, distinguishing between pneumonia and normal cases. Through testing on the dataset, LR-Net achieved an accuracy of 91.19%, with 268.098 K of parameters and 732.2184 M FLOPs.

The significant reduction in computing resources and memory requirements makes LR-Net an ideal candidate for deployment on mobile and edge computing devices [14].It provides healthcare practitioners with a scalable, accessible solution that improves diagnostic insight into pneumonia.In addition, the lightweight nature of LR-Net paves the way for rapid and accurate diagnosis of pneumonia in resource-limited settings, potentially reducing pressure on global healthcare systems and ensuring timely treatment for patients[15]. The main contributions of this paper are as follows:

(1)Application of residual network: We use residual nconnection to avoid the problem of feature degradation common in medical image processing. With this approach, our model is able to efficiently perform the binary classification task of pneumonia, which is essential to improve the accuracy of diagnosis.

(2)Lightweight model innovation: We have innovatively improved the traditional ResNet and developed a lightweight model structure. This structure significantly reduces the number of parameters for the model, enabling it to run on

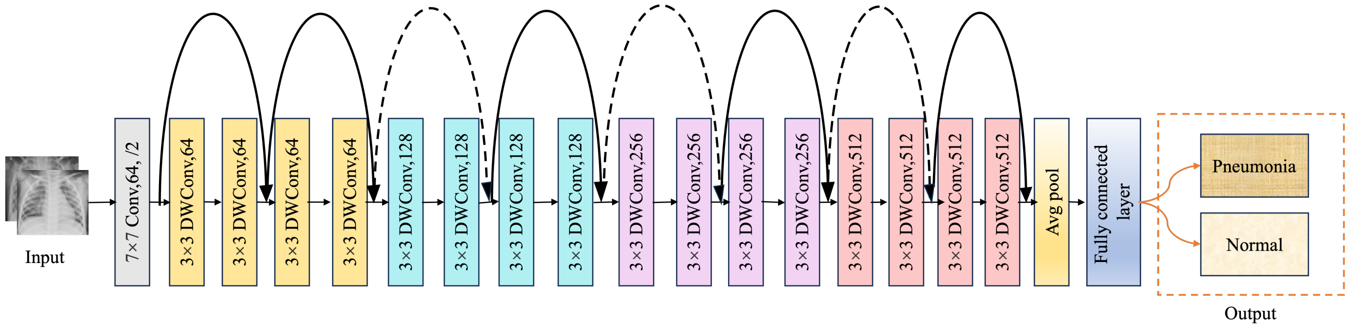

Fig. 1. LR-Net's structure

resource-constrained devices while ensuring accuracy and speed of diagnosis.

(3)Comparison with other CNNs[17][18][19] LR-Net shows superior performance compared to currently popular classical neural network models. Through a series of detailed experimental comparisons, LR-Net performed well on several key performance indicators including accuracy, accuracy, recall and F1-score. This result not only demonstrates the validity of our model, but also highlights its potential application in the field of medical image analysis.

The rest of paper is as follows: Section II introduces the overall structure and the use of related concept. Section III introduces the source of the dataset that we use, experimental environment setup, the calculation of evaluation indexes, LR-Net results, as well as the experimental results comparing LR-Net with other CNNs. Section IV presents the final conclusion.

II. Methods

In this study, we propose LR-Net, a lightweight deep learning network that combines the advantages of the ResNet architecture with the lightweight characteristics of DWConv, specifically for the automated diagnosis of pneumonia and its structure is shown in Fig. 1. The network aims to improve the accuracy of pneumonia detection by accurately identifying disease features in chest X-ray images, while significantly reducing the need for computing resources.

A. Overview of the proposed method

Our workflow begins with the pre-processing of the chest X-ray image, including the standardization and enhancement of the image, which is first cropped with random size and aspect ratio and then scaled to 224×224 pixels. A random horizontal flip is then performed, and finally the image is converted into a tensor and normalized with a preset mean and standard deviation, ensuring that the network can effectively extract features from a variety of image qualities. After that, the image is fed into the LR-Net, which consists of multiple lightweight residual blocks (LR-blocks), each consisting of a depth-separable convolution layer (DWConv) and residual connections.

Inside the LR-Net, each LR-Block processes input data through its depth-separable convolutional layers, which reduce the number of parameters through grouping operations while preserving continuity of information flow through residual connections. This structural design helps to reduce the complexity of the model while maintaining or enhancing the feature extraction capability. After continuous LR-Blocks processing, the model fuses features through the fully connected layer, and finally outputs a binary classification result to determine whether the image represents pneumonia or normal.The entire network structure is optimized for efficient forward propagation, ensuring rapid inference even in resource-constrained environments.

In the process of model training, we use the cross-entropy loss function to optimize the model parameters and verify the model performance through a series of experiments. Tests on chest X-ray image datasets show that our LR-Net performs well on several evaluation metrics including accuracy, accuracy, recall and F1-score. Compared with traditional deep learning models, LR-Net has higher computational efficiency while maintaining high diagnostic performance, and the number of parameters and computation amount are greatly reduced. Compared with traditional deep learning models, LR-Net has higher computational efficiency while maintaining high diagnostic performance, and the number of parameters and computation amount are greatly reduced.

B. Depthwise convolution (DWConv)

DWConv is an efficient convolutional method, which is widely used in modern convolutional neural networks to reduce the number of model parameters and computational costs, while maintaining model performance. In our proposed LR-Net, DWConv is one of the key components of a lightweight network.

The traditional convolution operation is to convolve all channels of the input feature map through a filter, and this process takes into account both spatial (pixels in the spatial dimension) and depth (different channels) information aggregation. In contrast, DWConv splits the process into two separate operations: deep convolution and point-by-point convolution.

In deep convolution, each input channel is convolved independently with a filter, meaning that each channel is only responsible for extracting its own features. This step greatly reduces the amount of computation because it no longer requires convolution across all input channels. Next, the point-by-point convolution combines the output of the deep convolution with a convolution kernel of 1×1, which allows the model to establish connections between different feature channels.

The main advantage of DWConv is its high efficiency. Due to the reduction in the number of parameters and the amount of computation, DWConv is particularly suitable for computing power constrained environments such as mobile devices and edge computing devices. In addition, compared to traditional convolution, DWConv can significantly increase the speed and reduce power consumption while maintaining the same network depth and complexity.

C. Residual block

Residual block is a revolutionary concept in deep learning, first proposed by He et al.[12] in his seminal ResNet paper. These blocks make it possible to train deeper neural networks because they solve the problem of disappearing gradients and exploding gradients by introducing residual connections. In the residual block, the input is transmitted not only through the convolutional layer, but also directly forward by skipping connections, allowing the gradient to flow directly through the network, which improves the efficiency and effectiveness of network training.

The core idea of the residual block is to let the network learn the residuals mapping based on the input, rather than directly learning the expected output. In this way, even if more layers are added, the network performance will at least not deteriorate, because these layers can learn the identity mapping so that the output equals the input.

D. Lightweight residual block (LR-Block)

In this study, we further innovatively design lightweight residual block (LR-Block), which significantly reduces the model parameters and computation while retaining the advantages of residual block .

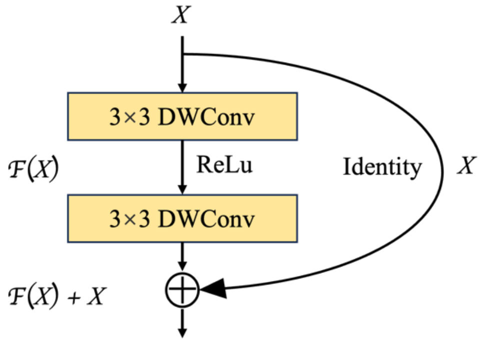

Fig. 2. LR-Block

As shown in the Fig. 2, the LR-Block consists of two main parts: a 3×3 DWConv layer and a residual connection. Each DWConv layer is followed by an activation function, ReLU, to introduce nonlinearity and improve the expressiveness of features.

In the design of the LR-Block, the input X first passes through a 3×3 DWConv layer, then through the ReLU activation function, and then into a second 3×3 DWConv layer. This tandem DWConv layer handles complex features and reduces the number of parameters while maintaining efficient feature learning. This design not only reduces the computational burden of the model, but also ensures the efficiency of feature transmission. After the DWConv layer, the input X is directly added to the nonlinear transformed feature by a short-circuit connection, an operation called residual learning. Finally, the nonlinear activation is completed by another ReLU function, and the output of the LR-Block is obtained.

The introduction of LR-Block makes LR-Net more efficient in deep feature learning, so that the model can maintain excellent performance even when the number of parameters is reduced.

III. Experiment

A. Dataset

To ensure that our research is based on high-quality medical images, we chose the recognized ChestX-ray2017 dataset [16] for our experiments. This dataset is recognized because it consists of chest X-ray images of actual patients, each of which has been screened by two specialists and corrected by a third to minimize diagnostic errors. This rigorous screening and calibration process provides us with a highly accurate and reliable data source, making our findings clinically relevant and credible.

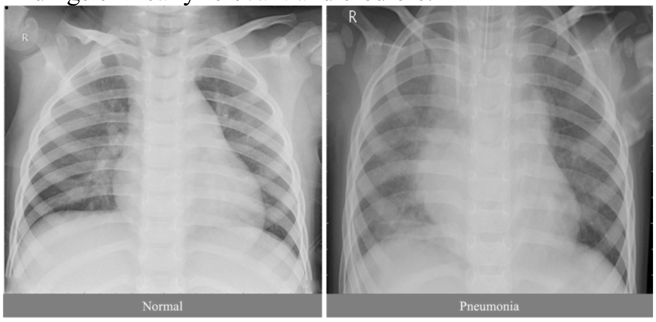

Fig. 3. Example images of our dataset

The dataset contained a total of 5857 chest X-ray images, which were divided into a training set and a testing set. Of these, the training data consisted of 5233 images, including 1349 images showing normal lungs and 3884 images showing characteristics of pneumonia. The testing set consisted of 624 images, including 234 normal images and 390 pneumonia images. This allocation ensures that we are able to train and validate our model on a relatively balanced data set, while covering a wide range of variations in health and disease states.

By using this dataset, our LR-Net is able to learn to distinguish the nuances of lung health to provide accurate pneumonia diagnosis in practical applications. This diversity of the dataset also supports the generalization ability of our model, which is not limited to images similar to the training data, but is able to perform well in never-before-seen, real-world situations with changing conditions.

B. Experimental setup

Our research experiments were performed in a local operating environment equipped with Intel Corei7 processor and NVIDIA RTX4090 graphics card, which combined with 32GB RAM ensured sufficient computing resources. The companion uses the Python3.8 environment and relies on scientific computing libraries such as PyTorch1.7.1 to build and train the LR-Net. During the experiment, we adopted Adam optimizer to optimize the model, set the learning rate to 0.00002, batchsize to 16, and trained 30 epochs to achieve the best performance of the model. To evaluate the model performance, we used the cross-entropy loss function and achieved a high degree of repeatability of the experiment through an automated Python script. We use the 5 fold cross-validation method to ensure the stability and reliability of the evaluation results. Finally, key indicators such as accuracy, recall, and F1-score were recorded to comprehensively evaluate the model's performance on the pneumonia classification task.

C. Evaluation metrics

In our study, we used standard binary evaluation indicators to evaluate the performance of our proposed LR-Net on the pneumonia classification task. Specifically, we use four indicators: true positive case (TP), false positive case (FP), true negative case (TN) and false negative case (FN) to comprehensively measure the diagnostic performance of the model. TP represents the number of pneumonia images correctly identified by the model. FP is the number of normal images incorrectly labeled by the model as pneumonia. TN is the number of normal images accurately predicted by the model. FN refers to pneumonia cases missed by the model. Based on these metrics, we calculated accuracy, precision, recall, and F1-score to quantify model performance. The accuracy reflects the overall correctness of the model. The recall measures the ability of the model to detect all relevant cases. The F1-score is a harmonic average of precision and recall, used to balance the relationship between the two, and is particularly useful for evaluating model performance on unbalanced datasets. These metrics ensure that we can rigorously evaluate LR-Net's performance in terms of medical diagnostic accuracy. The calculation formula is as follows:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1-score = \frac{2Precion \times Recall}{Precision+Recall} = \frac{2TP}{2TP+FN+FP} \tag{4}$$

D. Effectiveness of LR-Net

The core advantage of LR-Net is its ability to maintain a high level of accuracy while reducing model parameters and computational effort, which was confirmed in our experiments. And traditional ResNet effects,parameter number pairs such as TABLE I and TABLE II As shown:

TABLE I. EXPERIMENTAL RESULTS OF LR-NET AND RESNET18

Models	Accuracy	Precision	Recall	Kappa	F1-score
ResNet18	88.12%	85.05%	80.63%	65.07%	82.49%
LR-Net	91.19%	89.28%	88.21%	77.45%	88.72%

TABLE II. NUMBER OF PARAMETERS OF LR-NET AND RESNET18

Models	Params	FLOPs
ResNet18	11.2190M	27.3359G
LR-Net	268.0980K	732.2184M

In order to better intuitively feel the differences between the two models, we analyzed TABLE I And TABLE II, as shown in Fig. 4 and Fig. 5.

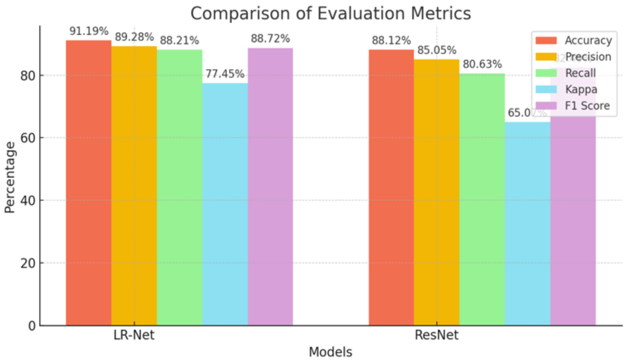

Fig. 4. Comparison of LR-Net and ResNet performance

Obviously, the accuracy of LR-Net reached 91.19%, which is better than the 88.12% of the traditional ResNet model, and this gap indicates the high efficiency of LR-Net in identifying pneumonia images.In terms of Precision, LR-Net also beat ResNet's 85.05% score with 89.28%, further demonstrating its reliability in distinguishing correct pneumonia cases from non-pneumonia cases. In addition, LR-Net's Recall of 88.21% was higher than ResNet's 80.63%, indicating that LR-Net was more comprehensive in catching pneumonia cases.

The kappa coefficient is another key measure of model consistency, and the kappa coefficient of LR-Net is 77.45%, compared to 65.07% for ResNet, showing better consistency and robustness. The F1-score is the harmonic average of accuracy and recall, and LR-Net achieves 88.72% in this indicator, while ResNet is 82.49%, further strengthening the performance superiority of LR-Net.

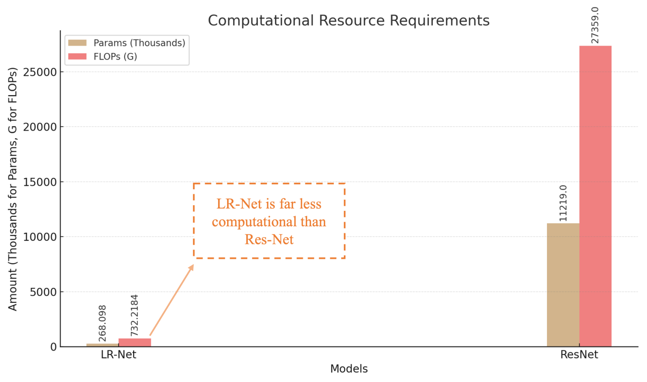

Fig. 5. Comparison of LR-Net and ResNet computations

As shown in Fig. 5, the table of LR-Net is particularly prominent in terms of the number of model parameters and the amount of calculation. With only 268.098 K parameters and 732.2184 M FLOPs, the lightweight design of the LR-Net significantly reduces resource consumption compared to ResNet's 11.2190 M parameters and 27.3359 G FLOPs.

E. Comparing with classification CNNs

Then we compare the LR-Net model with other CNNs. The experimental results are shown in TABLE III and TABLE IV:

TABLE III. EXPERIMENTAL RESULTS OF LR-NET AND OTHER NETS

Models	Accuracy	Precision	Recall	Kappa	F1-score
LeNet	86.11%	85.24%	77.26%	60.37%	80.00%
MobileNet V2	87.07%	83.04%	89.24%	70.19%	84.93%
AlexNet	89.18%	84.77%	88.29%	72.58%	86.26%
VGG16	89.46%	85.54%	89.32%	74.24%	87.09%
LR-Net	91.19%	89.28%	88.21%	77.45%	88.72%

TABLE IV. NUMBER OF PARAMETERS OF LR-NET AND OTHER NETS

Models	Params	FLOPs
LeNet	5.6120M	56.2M
MobileNetV2	3.1981M	39.9G
AlexNet	50.8440M	1.01G
VGG16	134.2687M	15.47G
LR-Net	268.0980K	732.2M

In order to get a better sense of the differences between the different models, let's compare TABLE III and TABLE IV visualization was performed, as shown in Fig.6 and Fig.7:

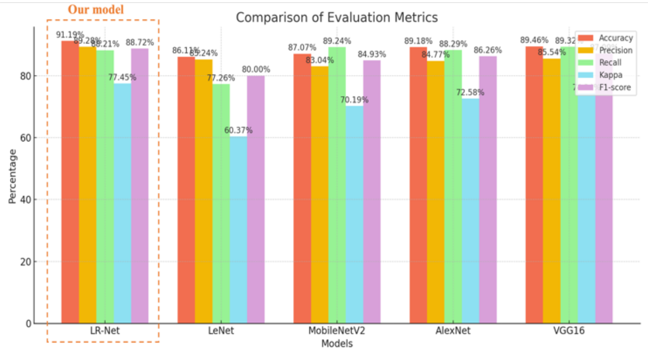

Fig. 6. Comparison of LR-Net and other CNNs performance

As shown in Fig.6, the accuracy of LR-Net reached 91.19%, higher than that of all comparison models, which proved its excellent performance in the overall classification correctness. In terms of precision (89.28%) and recall (88.21%), LR-Net also demonstrated its strong ability to reduce false positives and correctly identify true positives. In addition, LR-Net had the highest Kappa coefficient (77.45%) and F1-score (88.72%), further underscoring its superiority in ensuring diagnostic consistency and balancing accuracy with recall.

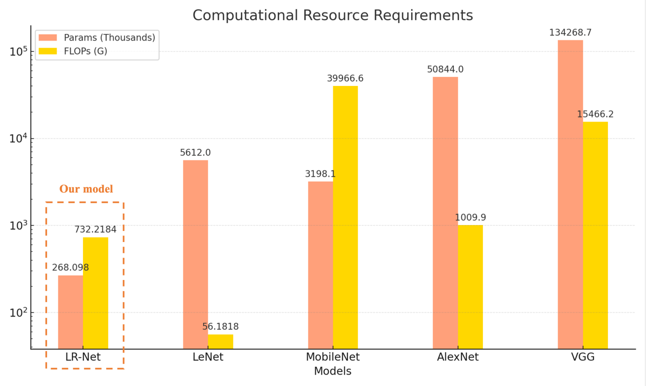

Fig. 7. Comparison of LR-Net and other Nets computations

In terms of the number of parameters in the model, the lightweight advantage of LR-Net is still obvious. Other models such as LeNet, MobileNet, AlexNet, and VGG have millions to billions of parameters and FLOPs respectively, which not only increases the storage and computing cost, but also slows down the reasoning speed of the model.

IV. CONLUSION

This study constructs LR-Net, an innovative CNN that has demonstrated excellent performance in early detection of pneumonia from chest X-ray images. Our experimental results prove that LR-Net has not only high accuracy, but also high parameter efficiency and low calculation cost, which has good practical value. The lightweight architecture of the model does not compromise its diagnostic integrity, ensuring that LR-Net can be effectively applied in a variety of clinical settings, including those with limited computing power. It paves the way for more convenient, efficient and reliable diagnosis of pneumonia, with the potential to reduce pressure on global healthcare systems and provide patients with timely interventions.

Taken together, LR-Net is expected to provide a more flexible and cost-effective approach to pneumonia detection. Future studies may focus on expanding the model's applicability to other medical imaging tasks and integrating it into real-world clinical workflows, where its impact on patient outcomes can be further assessed.

REFERENCES

[1] Cunha, B. A. (2006). The atypical pneumonias: clinical diagnosis and importance. *Clinical Microbiology and Infection*, *12*, 12-24.

[2] Wang, S. X., Wang, Y., Lu, Y. B., Li, J. Y., Song, Y. J., Nyamgerelt, M., & Wang, X. X. (2020). Diagnosis and treatment of novel coronavirus pneumonia based on the theory of traditional Chinese medicine. *Journal of integrative medicine*, *18*(4), 275-283.

[3] Zhou, W., Wang, H., & Wan, Z. (2022). Ore image classification based on improved CNN. *Computers and Electrical Engineering*, *99*, 107819.

[4] Taspinar, Y. S., Cinar, I., & Koklu, M. (2022). Classification by a stacking model using CNN features for COVID-19 infection diagnosis. *Journal of X-ray science and technology*, *30*(1), 73-88.

[5] Vankdothu, R., Hameed, M. A., & Fatima, H. (2022). A brain tumor identification and classification using deep learning based on CNN-LSTM method. *Computers and Electrical Engineering*, *101*, 107960.

[6] İnik, Ö. (2023). CNN hyper-parameter optimization for environmental sound classification. *Applied Acoustics*, *202*, 109168.

[7] X. Wu *et al.*, "CTransCNN: Combining transformer and CNN in multilabel medical image classification," *Knowledge-Based Systems,* vol. 281, p. 111030, 2023/12/03/ 2023.

[8] Thakur, S., & Kumar, A. (2021). X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN). *Biomedical Signal Processing and Control*, *69*, 102920.

[9] Xu, Y., Lam, H. K., Jia, G., Jiang, J., Liao, J., & Bao, X. (2023). Improving COVID-19 CT classification of CNNs by learning parameter-efficient representation. *Computers in Biology and Medicine*, *152*, 106417.

[10] Tan, W., Liu, P., Li, X., Liu, Y., Zhou, Q., Chen, C., ... & Zhang, Y. (2021). Classification of COVID-19 pneumonia from chest CT images based on reconstructed super-resolution images and VGG neural network. *Health Information Science and Systems*, *9*, 1-12.

[11] Chua, L. O. (1998). *CNN: A paradigm for complexity* (Vol. 31). World Scientific.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[13] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

[14] Sun, X., & Ansari, N. (2016). EdgeIoT: Mobile edge computing for the Internet of Things. *IEEE Communications Magazine*, *54*(12), 22-29.

[15] Blozik, E., Wildeisen, I. E., Fueglistaler, P., & von Overbeck, J. (2012). Telemedicine can help to ensure that patients receive timely medical care. *Journal of telemedicine and telecare*, *18*(2), 119-121.

[16] Kermanyet DS, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–31.

[17] El-Sawy, A., El-Bakry, H., & Loey, M. (2017). CNN for handwritten arabic digits recognition based on LeNet-5. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2* (pp. 566-575). Springer International Publishing.

[18] Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(10), 143-150.

[19] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).