Pneumonia X-ray Image Classification Methods Based on Deep Learning

Xinyu Zhou

College of Software, Taiyuan University of Technology, Taiyuan, Shanxi, China
2023006257@link.tyut.edu.cn

Abstract. Pneumonia is a respiratory disease that is widely present in daily life. For pneumonia, timely and accurate diagnosis is very crucial for clinical treatment. The traditional chest X-ray diagnostic method mainly relies on all the professional knowledge of radiologists to carry out the diagnostic work. This diagnostic method is time-consuming and energy-consuming. In the diagnostic process, it is also prone to the subjectivity of the doctor himself. To enhance the effectiveness and accuracy of diagnosis, this paper employs a publicly available chest X-ray dataset to construct and train three deep learning models with different architectures. These models include custom-designed deep convolutional neural networks and transfer learning models based on ResNet50. And an integrated model that integrates the features of CNN and VGGNet. The results obtained from the experiments show that the ensemble model outperforms other models in all evaluation criteria, achieving relatively high accuracy, precision, recall rate and F1 score. This paper confirms that deep learning and ensemble learning strategies are effective in medical image analysis and have great potential. It also provides valuable technical guidance and practical foundations for the development of efficient computer-aided diagnostic tools.

Keywords: Pneumonia Classification, Medical Image Evaluation, Convolutional Neural Network, Transfer Learning, Ensemble Learning

1 Introduction

Pneumonia is one of the most common infectious diseases worldwide. It poses very serious health risks to people. For children and the elderly, pneumonia is also associated with an increased incidence and mortality rate [1]. In the field of clinical medicine, chest X-rays are the most commonly used method for detecting pneumonia. It is also a relatively economically feasible imaging method.

However, when analyzing X-ray images, it largely depends on the proficiency of radiologists and the practical knowledge they possess. The ever-increasing volume of image data imposes a heavy burden on radiologists. Additionally, visual fatigue and the subjective assessment among different doctors may lead to inconsistent diagnostic results. This inconsistency in diagnostic results may even lead to the neglect of pneumonia cases or the occurrence of misdiagnosis [2].

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_60

In the past several years, as artificial intelligence, or AI technology, has advanced at a breakneck pace, deep learning, which encompasses convolutional neural networks, has spurred revolutionary growth and advancement in the domain of computer visualization. It possesses a robust capacity for automatic feature extraction and learning, which offers a highly significant outlook for medical image analysis tasks. These tasks in medical image analysis span areas like tumor identification, lesion delineation, and disease categorization[3].

The primary objective of this research is to explore and contrast the effectiveness of various deep - learning frameworks during the binary categorization of pneumonia X - ray images, with the aim of finding an effective and accurate automatic diagnosis method.

To achieve this goal, this study first selected a publicly available and annotated pediatric chest X-ray dataset, and finally decided to use this dataset. After a comprehensive analysis of the data, a series of preprocessing steps were carried out. These steps included normalizing the image size, normalizing the pixel values, increasing the data volume, and addressing the issue of class imbalance.

Then, the research implemented three deep learning models with different levels of complexity and strategies. The first one is the self-constructed deep convolutional neural network, namely deep CNN, which is used to evaluate the performance of the basic CNN architecture. The second one is the transfer learning model, namely ResNet_Transfer, which uses the pre-trained ResNet50 network to test whether transfer learning is effective in medical imaging tasks. The third one is an innovative ensemble learning model, namely the ensemble model. This model combines the custom-designed CNN branch and the feature extraction branch based on VGG16. This combination takes advantage of the respective strengths of the two models and can capture more different types of image features.

By training and evaluating these three models on the same dataset, systematically comparing their performance differences and analyzing the root causes of these differences, this study aims to provide practical evidence for the central theme, which is "developing a reliable and efficient deep learning-based automatic classification system for pneumonia X-ray images".

2 Dataset and Methods

2.1 Dataset

The data used in this study was obtained from the publicly available dataset "Chest X-ray Images", which contains 5,863 chest X-ray images of children aged 1 to 5. These images were from Guangzhou Women and Children's Medical Center. Before using these images for model training, all the images were screened by professional doctors. And annotations were also made.

The dataset is divided into three parts, including the training subset, the test subset and the validation subset. It contains two types of situations, one is "pneumonia" and the other is "normal". The distribution of the initial dataset has a very prominent class

imbalance problem. To elaborate, pneumonia samples account for approximately 74.3% of the entire dataset. Such a situation poses a challenge to model training.

To enhance the performance of the model and improve its generalization ability, the following preprocessing work was carried out on the data in this study:

First is Image normalization. Preparing a high-quality dataset through pathological image preprocessing methods is an effective way to improve the diagnostic accuracy of the model [4]. In this study, all input images are resized to the fixed size required by the model input (e.g., 224x224 pixels) and pixel values are normalized from the range [0, 255] to the range [0, 1] to accelerate model convergence.

Second is Data Augmentation. To expand the training dataset and alleviate the problem of overfitting, this study applies a variety of online data augmentation techniques to the training images, including random rotation (±15 degrees), random horizontal translation (±10%), random scaling (±20%), and random horizontal flipping.

Third is Dataset Repartitioning. Considering the relatively small number of validation set samples in the original dataset, the initial training set is divided into a fresh training set and a validation set in this research, with the proportion being 80% for the new training set and 20% for the validation set, to obtain a more reliable model performance evaluation.

Fourth is Class Weight Balancing. To address the data imbalance problem, class weight balancing is introduced during model training. By assigning higher weights to the "normal" category with fewer samples, the model can pay more attention to the samples of the minority class during training, avoiding the model being biased towards predicting the majority class.

2.2 Model Architecture

Deep CNN Model. Convolutional neural networks are a type of deep feedforward neural network. They feature local connectivity, weight sharing, and hierarchical representation. In many computer vision-related works, convolutional neural networks have achieved particularly outstanding results. They are also a widely used model in the field of deep learning [5].

The network architecture mainly consists of the following components. Firstly, there are four convolutional modules, each of which is composed of one or more convolutional layers, batch normalization layers, and Max pooling layers. As the network deepens continuously, the number of convolutional kernels also keeps increasing, with the numbers being 32, 64, 128, and 256 respectively. Such a setting can help the model gradually learn the relevant functions. Moreover, for the purpose of regularization, a layer that implements dropout is inserted prior to the fully connected layer. This layer randomly eliminates neurons with a specific probability to reduce the interdependence between neurons. Moreover, for classifier, two fully connected layers (Dense) are connected on top of the convolutional base, with 512 and 256 neurons respectively, and finally the probability of the sample belonging to the pneumonia category is obtained through an output layer with a Sigmoid activation function.

ResNet50 Transfer Learning Model. The ResNet model effectively solves the problem of model degradation in deep networks through residual blocks and skip connections, improving training stability and model expressiveness [6]. This study uses the classic ResNet50 model as the base network.

Base network: Load the ResNet50 model pre-trained on the ImageNet dataset and freeze the weights of most of its convolutional layers. These bottom and middle convolutional layers have learned rich common visual features (such as edges, textures, shapes, etc.) and can be directly used for new tasks.

Custom classifier: Freeze the original bottom-level feature extractor and top-level classifier of ResNet50, and add a new classifier adapted to the binary classification task of this study. This classifier usually consists of a global average pooling layer (GlobalAveragePooling2D), a fully connected layer, and a Sigmoid output layer.

Fine-tuning: After the initial training phase, the last few layers of the pre-trained model can be selectively unfrozen (the last 20 layers are selected in this system) and fine-tuned using a very small learning rate. This enables the model to optimize high - order features so as to more effectively conform to the distinctive patterns present in X - ray images while retaining general features.

Ensemble Model. Ensemble learning effectively improves the overall accuracy and robustness of predictions by integrating the advantages of multiple base models [7]. To integrate the advantages of diverse models, this study designed a two-branch ensemble model aimed at simultaneously capturing local detail features and global semantic information of images.

The model structure is shown in Figure 1. It contains two parallel feature extraction branches: one is a lightweight custom CNN network similar to that described in Section 2.2.1, used to extract local and fundamental features of the image; the other is a feature extractor based on a VGG16 pre-trained model [8]. Similar to ResNet50, its pre-trained weights on ImageNet are loaded and frozen. VGG16 is known for its concise yet deep structure, effectively extracting powerful semantic features. Once the branches are established, the feature maps retrieved from the two branches will be combined along the channel axis. This method can effectively fuse the different features obtained by different models together, creating a more comprehensive and distinguishable feature profile. Then, the merged features will be input into the shared classifier. These features will pass through the fully connected layer and the S-shaped activation function before the classification result can be obtained.

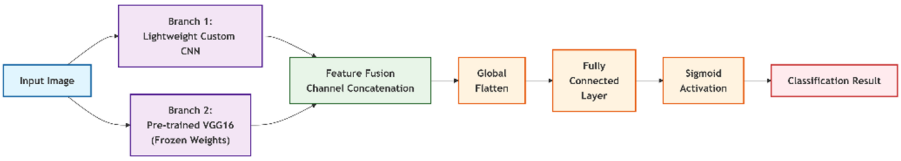

Fig. 1. schematic diagram of the structure of the two-branch ensemble model. (Picture credit: Original)

3 Experiment

3.1 Experimental Configuration

This experiment was conducted on a computer equipped with a standard CPU training setup. The deep learning framework used was TensorFlow, with a version of approximately 2.20.0. It was also supplemented by some Python libraries. Libraries like Matplotlib, Seaborn, Scikit-learn, NumPy and Pillow. The function of these libraries is to conduct data processing, model evaluation and result visualization.

During the training phase, all models selected the Adam optimizer. The initial learning rate was set to 0.001, the batch size was set to 32, and the upper limit of training was 25 epochs. The specific details of the experimental configuration can be seen in Table 1.

To alleviate overfitting and enhance training efficiency, this algorithm introduces two callback functions, namely early stop and learning rate scheduling. The early stop feature tracks the validation loss. If the validation loss does not improve after five consecutive attempts, the training will be stopped in advance [9]. The choice of learning rate has a significant impact on model performance when training convolutional neural networks [10].

Table 1. Experimental configuration details.

Configuration Category	Specific Item	Parameters/Description
Hardware Environment	Computer	Standard CPU Training Environment
Software Environment	Deep Learning Framework	TensorFlow（~2.20.0）
	Key Python Libraries	Matplotlib, Sea born, Scikit-learn, Numpy, Pillow
Training Hyperparameters	Optimizer	Adam
	Initial Learning Rate	0.001
	Batch size	32
	maximum training epochs	25
callback function	early stopping mechanism	monitoring metric: val_loss Patience value: 5
	Learning rate scheduling	monitoring metric: val_loss Function: Automatically reduce the learning rate when loss stagnates

3.2 Experimental Results and Analysis

Model Performance Comparison. Table 2 summarizes the evaluation metrics of the three models. All exceeded the 65% baseline, showing strong performance. The ensemble model performed best on all key metrics except precision, achieving the highest recall (0.8504 vs. 0.8034 for Deep CNN and 0.8076 for ResNet_Transfer), indicating superior ability in identifying pneumonia cases—crucial for reducing missed diagnoses in clinical screening. Although ResNet_Transfer achieved slightly higher accuracy, it also showed a higher false positive rate. Overall, the ensemble model demonstrated the best comprehensive performance.

Table 2. Comparison of the performance of the three models on the test set.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Deep CNN	0.8432	0.7737	0.8034	0.7882	0.9192
Res Net Transfer	0.8846	0.8750	0.8076	0.8400	0.9443
Ensemble Model	0.8958	0.8690	0.8504	0.8596	0.9536

Training Process Analysis. During the entire training process, the accuracy and loss curves of the three models can be seen in Figures 2, 3 and 4.

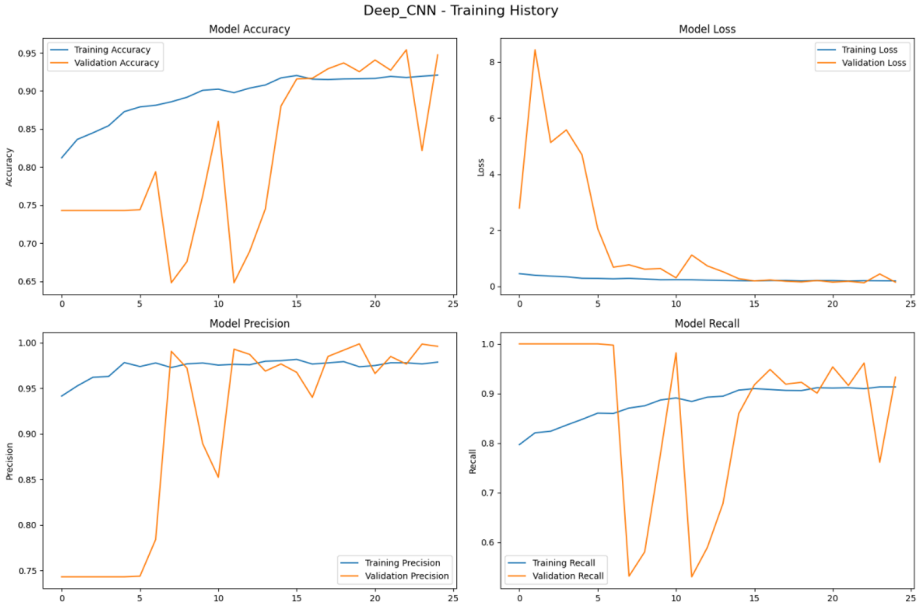

Fig. 2. Deep CNN model training curve. (Picture credit: Original)

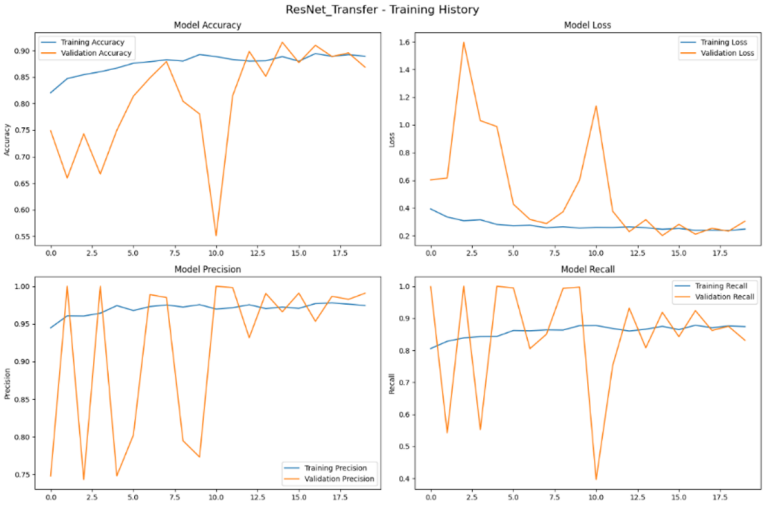

Fig. 3. ResNet_Transfer learning model training curve. (Picture credit: Original)

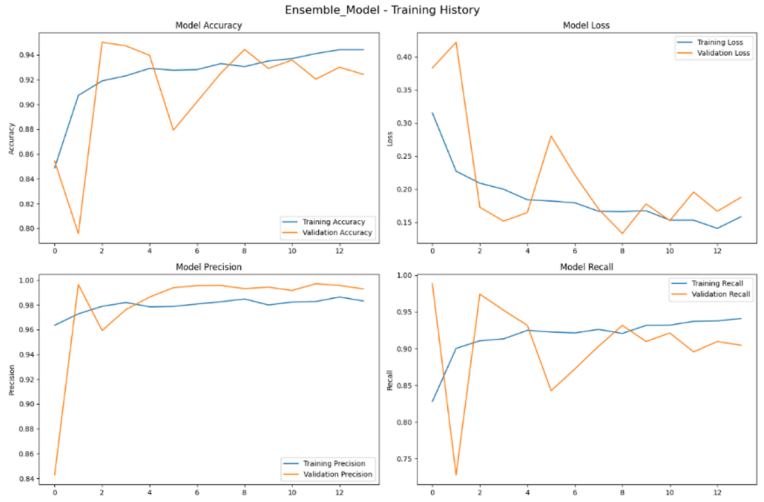

Fig. 4. Ensemble learning model training curve. (Picture credit: Original)

After conducting a comparative study on three curve sets, it was found that all models show a good convergence trend, with the accuracy gradually increasing and the loss rate steadily decreasing. It can be guaranteed that there is sufficient model training.

The Ensemble_Model has obtained the best training conditions. Its validation metric curve is very smooth and very close to the training curve. There is only a small gap between these two curves. This model has excellent generalization ability and elasticity against overfitting. It is a very reliable learning process.

Overall, Deep_CNN shows a relatively stable training situation and has also carried out validation work. The trend of the validation curve is generally consistent with that

of the training curve. However, during the time period from the 5th to the 10th round, there will be some fluctuations. When the model is learning in the early stage, it is relatively sensitive to data features. However, effective convergence was achieved in the end.

ResNet_Transfer shows relatively weak stability. Although its final performance is quite good, its validation loss curve experiences a relatively large peak in the later stage of training, while the accuracy and recall curves show significant fluctuations. This might be due to the problem of adaptation conflicts between the general features obtained during the pre-training process and the specific features of the current medical images, which makes the model less stable during the adaptation process.

For the analysis of overfitting, due to the adoption of strategies such as data improvement, Dropout regularization, and early stopping, none of the three models exhibited severe overfitting. The difference between the training curve and the validation curve of the Ensemble_Model was the smallest, which proved that it had the optimal overfitting control effect. Overall, a visual analysis of the training process verified that all models achieved a relatively high performance level. It also pointed out that the ensemble learning strategy has a very prominent advantage in improving the stability and robustness of model training.

Confusion Matrix And Error Evaluation. As can be seen from Figure 5, the confusion matrix can present the specific prediction situation of the model for each category in the test set in an intuitive way.

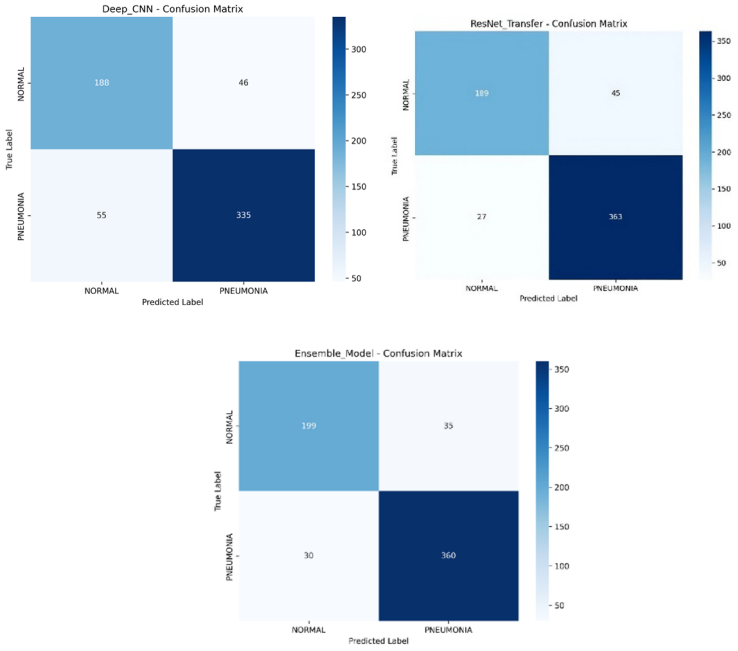

Fig. 5. Confusion Matrix for the models. (Picture credit: Original)

After conducting a comparative analysis of the model prediction results in this paper, the following conclusions are drawn:

The Ensemble_Model shows the best performance in false negative control, with only 35 cases, indicating the lowest rate of misclassified pneumonia patients—consistent with its highest recall in Table 1. Although it produces slightly more false positives (30) than ResNet_Transfer (27), the level remains acceptable for medical screening, where minimizing missed diagnoses is prioritized. ResNet_Transfer achieves the highest accuracy by better suppressing false positives, while the Ensemble_Model offers a more balanced trade-off, demonstrating superior overall discriminative capability.

ROC Curve Analysis. As shown in Figure 6, the receiver operating characteristic curve can illustrate the correlation between the true positive rate and the false positive rate of the classification model under different thresholds, and the area under the curve is a very crucial indicator for measuring the overall effectiveness of the model.

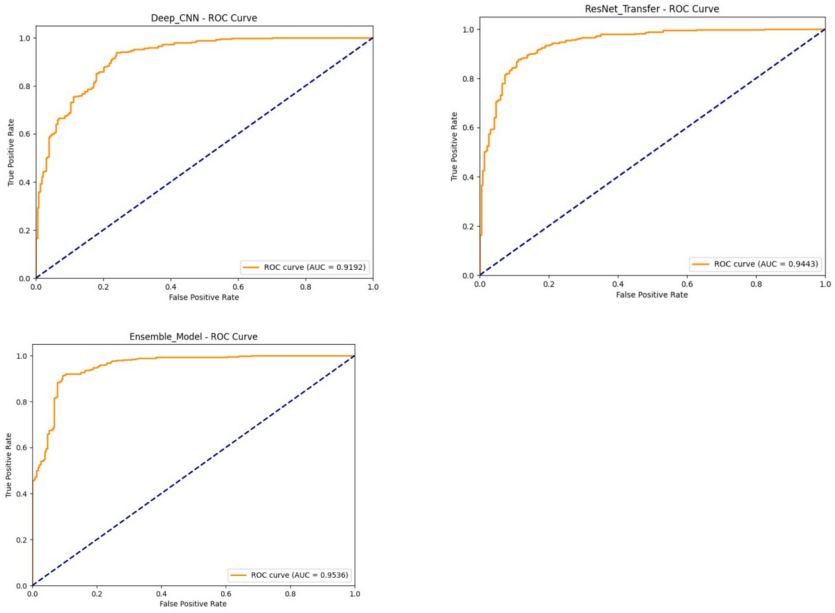

Fig. 6. ROC Curve of the Models. (Picture credit: Original)

All models achieved AUC values above 0.91, indicating strong discrimination between positive and negative cases. The ensemble model performed best, with an AUC of 0.9536 and an ROC curve closest to the upper-left corner, confirming its superiority. This performance stems from its dual-branch architecture: a custom CNN branch capturing fine-grained textures and a VGG16-based transfer branch providing rich semantic information. Their feature fusion yields more comprehensive representations, enhancing robustness and accuracy.

Compared with models trained from scratch, the ResNet_Transfer model benefits from pre-trained weights, validating the effectiveness of transfer learning in limited-data medical imaging tasks [11]. Clinically, high recall is essential to minimize missed pneumonia diagnoses [12]. The proposed ensemble model meets this need with a recall of 0.8504, slightly sacrificing accuracy but greatly reducing false negatives. In practice, all predicted positives can be rechecked by radiologists, enabling effective human–machine collaboration and reducing overall diagnostic risk.

4 Conclusion

This paper develops three deep learning models, which are used to automatically classify pneumonia from chest X-rays. These models have been trained and evaluated on public datasets, and all models have achieved very good performance. Among them, the integrated model that combines custom CNN and VGG16 transfer learning In terms of accuracy, recall rate, F1-score and AUC, it performs better than the independent CNN and ResNet50 models. This clearly demonstrates the benefits that multi-feature fusion can bring.

However, this research also has some limitations. It is dependent on a single dataset, which may limit the model's generalization ability across institutions or different device types. Although there are weights, the imbalance among classes may still have an impact on the learning of a few classes. The integrated model is rather complex, which has raised concerns about inference speed and hardware requirements. Inference speed and hardware requirements are crucial for clinical deployment.

Future work should focus on three aspects: To enhance the robustness of the model and improve its generalization ability, this paper can perform integration operations on multi-center data. It can also adopt model compression methods such as pruning or knowledge distillation. These techniques can improve the efficiency of model deployment and keep the accuracy of the model at the original level without any decline. In addition, interpretable tools like Grad - CAM can be used to examine the basis on which decisions are based. Adopting this approach can enhance the level of trust in the clinical aspect, making it easier for doctors to use this model.

References

1. Zhou, X., Hong, H., Fang, T., et al.: Advances in Epidemiological Research on Pneumonia. Preventive Medicine 35(08), 682–686 (2023)
2. Xu, Y., Zheng, Z.: Analysis of the Evaluation and Influencing Factors of Radiology Imaging Diagnostic Reports. Education and Teaching Forum (03), 13–16 (2022)
3. Zhao, Y.: Research on Three-Dimensional Medical Image Segmentation Based on Deep Learning. Ph.D. thesis, University of Chinese Academy of Sciences, School of Engineering Sciences (2024)
4. Xie, P.: Research on Intelligent Diagnosis Methods for Skin Pathological Images Based on Deep Learning. Ph.D. thesis, National University of Defense Technology (2022)

5. Li, Y.: Research on Auxiliary Diagnosis Algorithms for Pulmonary Diseases Based on Deep Learning. Master's thesis, Yantai University (2024)
6. Fu, Z., Li, R., Jing, Y., et al.: Meibomian Gland Image Grading Method Based on Improved ResNet and Multi-Feature Fusion. Computer and Modernization, 1–11 (2025)
7. Wang, N., Wu, F., Zhao, Y., et al.: Text-Image Sentiment Analysis Method Based on Ensemble Learning and Multimodal Large Language Models. Computer Engineering and Applications, 1–11 (2025)
8. Jiang, Y., Ding, S., Wu, P.: Research on Multimodal Information Feature Classification Based on BiLSTM-VGG16. Information Studies: Theory & Application 44(11), 180–186+179 (2021)
9. Yu, Y., Luo, Y., Miao, J., et al.: Prediction of Unsteady Flow Fields around Aircraft Wings Based on Deep Learning and Influence Function Method. Missiles and Space Vehicles Technology (Chinese and English Edition) (05), 32–37 (2023)
10. Chen, L.: Research on Breast Cancer Pathological Image Classification Based on Convolutional Neural Networks. Master's thesis, Northwest Normal University (2020)
11. Tian, T.: Research on Lung CT Image Analysis Methods Based on Deep Learning. Master's thesis, Qilu University of Technology (2025)
12. Xie, X.: Research on Medical Image Segmentation Methods Based on Deep Learning. Master's thesis, Dalian Maritime University (2024)