Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33977

CNN-Transformer Hybrid Model for Accurate Pneumonia Detection in Chest X-Ray Images

Ruihan Chen

Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong, China
u3644921@connect.hku.hk

Abstract. This research proposes a hybrid deep learning framework that unites Convolutional Neural Networks (CNN) and Transformers for the check-up of pneumonia on chest X-ray images. Traditional CNNs have excellent ability in capturing high-resolution local space features, such as tiny changes in the lung parenchyma's texture, but they fail to capture long-distance context dependencies. To address the issue, the model possesses a Transformer component that utilizes self-attention mechanisms for modelling the global relations among lesions, which are scattered and anatomical structures. This research used a 5,856-image dataset (pneumonia and normal), which used complex preprocessing methods like median filtering and geometry expansions, to ensure strong generalization ability. The experiment measured result displays that the hybrid model does much better than separate CNNs and achieves 94.38% sensitivity and 89.96% accuracy. The analysis of a detailed confusion matrix can show that this architecture has an effect in reducing false negative cases, and it directly responds to the clinical demand that screening must have high sensitivity. These results point out that the model can act as a dependable instrument for diagnosis. It builds a bridge over medical knowledge gaps in scarce resource regions, and it can promote clinical workflow in an effective way.

Keywords: transformer, chest X-ray, pneumonia detection, deep learning

1. Introduction

Recently, artificial intelligence (AI) has promoted the growth of medical diagnostics, especially when its use is put in imaging technologies. Chest X-ray is an important tool for diagnosing respiratory diseases such as pneumonia, which is one of the main reasons for death all over the world. According to the World Health Organization, pneumonia brings unequal harm to children and old people alike. Pneumonia caused the death of over 808,000 children below the age of 5 in 2017, it holds 15% of all deaths of children who are under 5 years [1]. The crowds who are at risk of pneumonia also include elderly people who are older than 65 years and people who already have existing health issues. In China, radiologists have a heavy workload in reading X-ray images every day. They must undertake these burdensome viewing tasks, which require attention and patience. Reducing subjective human errors and improving efficiency have long been targets that everyone in this field would like to achieve [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33977

Deep learning machine learning techniques on chest X-rays are gaining popularity as they can be easily used with low-cost imaging techniques, and there is an abundance of data available for training different machine-learning models [3]. As the foundation stone of this area, CNNs have obtained very notable achievements in image classification and are now widely used by all research circles. These models often have better performance than traditional diagnostic methods on both speed and accuracy, and they become the necessary tools for current clinical use.

CNN has difficulty in capturing the relative positions of many image features; they need large receptive fields to code their combinations and therefore deal with long-distance dependencies. Another limitation is the use of large convolutional kernels, which demand greater computational resources to process [4]. According to these restrictions, this research put forward a Transformer-CNN model for promoting detection effect and decreasing model complexity. This article has as its objective the design of a hybrid model, which can promote the classification effect of AI on chest X-rays. The target is to develop an accurate diagnostic tool to enhance the efficiency of medical imaging analysis, relieve the burden of radiologists, and provide reliable results and support for patients.

2. Methods

2.1. Dataset description and preprocessing

2.1.1. Dataset description

The chest X-ray dataset for pneumonia detection used in this study is retrieved from the publicly available Kaggle repository, originally contributed by Madhav Mathur [5]. Although the dataset provides an official train/val/test split, this research re-divided it into training, validation, and test sets in a stratified 60:15:25 ratio to preserve the class distribution and achieve a more statistically reliable evaluation. A larger test set (25%) ensures a statistically robust evaluation of model generalization, which is critical for clinical reliability. The validation set (15%) supports stable hyperparameter tuning and early stopping. The remaining 60% is used for training to ensure sufficient feature learning under class imbalance. The total dataset contains 1583 normal chest X-ray images and 4273 pneumonia chest X-ray images, showing a class imbalance. The ratio between normal and pneumonia is approximately 1:3.2. The exact data is shown in Table 1.

Table 1. Dataset split and sample distribution

Data split	Normal	Pneumonia	Total
Training set (60%)	950	2564	3514
Validation set (15%)	237	641	878
Testing set (25%)	396	1068	1464
Total	1583	4273	5856

The images were imported and handled in advance by OpenCV (cv2) and NumPy libraries, with strict effectiveness checks for guaranteeing the data quality.

2.1.2. Preprocessing process

Initially, images are changed in size in the same way to 150×150 pixels, to satisfy the input requirement. Unreadable and broken images that have been filtered out are automatically processed

DOI: 10.54254/2755-2721/2026.33977

to guarantee image validity and enhance the efficiency of subsequent training after checking the directory and the paths that lead to the datasets.

To further image quality optimization, chest X-ray images are loaded in single-channel grayscale form instead of RGB, because grayscale images can retain enough pathological information for pneumonia detection, at the same time, they reduce the number of parameters that the model is required to learn. Grayscale images keep enough structural and intensity-relevant information for effective pixel-level similarity research, although color channels do not exist [6]. After resizing and grayscale conversion, pixel intensity values are scaled from [0, 255] to [0, 1] to standardize the input data distribution. Additionally, the grayscale image array is reshaped to add a channel dimension, which aligns with the input tensor requirement of CNNs.

2.1.3. Image enhancement

To improve the training stability and classification accuracy of the model, targeted image enhancement is applied to the training set. A mild and clinically reasonable data augmentation pipeline is adopted to avoid destroying pathological structures in chest X-ray images. Data augmentation has been reported to improve deep learning classification performance by augmenting existing data rather than collecting new data, can significantly increase the diversity of training data, and is crucial when the dataset is imbalanced [7]. In the training process, the image generator carries out small-angle random rotation that may reach 15 degrees, moderate random zoom with a ratio of 0.15, slight horizontal and vertical displacements that are within 10 percent of the image size, gentle shear transformation with a range of 0.05, and random horizontal flip operation. In order to maintain real lesion characteristics and anatomical structure, all transformations are maintained within a small scope. This assists the model in learning more firm visual expression representations without adding noise or distortion that is not real.

2.2. Model design

2.2.1. Pure CNN model construction

To extract discriminative local spatial features of pulmonary lesions from chest X-ray images, A lightweight yet effective five-layer convolutional neural network (CNN) architecture that balances feature-extraction capability with computational efficiency is designed in this research. The proposed model captures complex patterns and spatial dependencies across multiple convolutional layers, enabling efficient prediction and supporting clinicians in critical diagnostic decision-making [8]. The system outputs the probability of pneumonia-positive and pneumonia-negative classifications for each input image. The architecture is summarized in Table 2.

DOI: 10.54254/2755-2721/2026.33977

Table 2. Pure CNN model architecture

Stage	Structure	Feature
1	32×3×3 Conv → BatchNorm → 2×2 MaxPool (strides=2)	Accepts 150×150 single-channel grayscale chest X-ray images, extracts low-level features, and reduces dimensionality via pooling to lower computational cost.
2	64×3×3 Conv → Dropout (0.1) → BatchNorm → 2×2 MaxPool (strides=2)	Learns mid-level features; Dropout suppresses overfitting, and batch normalization stabilizes training.
3	64×3×3 Conv → BatchNorm → 2×2 MaxPool (strides=2)	Enhances mid-level feature representation to capture local contextual features of lung regions; no Dropout retains more feature information.
4	128×3×3 Conv → Dropout (0.2) → BatchNorm → 2×2 MaxPool (strides=2)	Extracts high-level pathological features; increased Dropout strength improves generalization.
5	256×3×3 Conv → Dropout (0.2) → BatchNorm → 2×2 MaxPool (strides=2)	Captures complex pneumonia-specific features; high channel count enhances feature discriminability.

The construction design follows a layered character study method in which the filter counts grow step by step from 32 to 256 through five different layers. This development permits the model to complete the transformation from capturing low-level visual features, such as the basic outlines of the lungs, to getting abstract and clinically meaningful descriptions in the deeper layers. These deeper layers put their focus on the form of disease damage, changes in density, and special pathological marks that are connected to pneumonia.

This structural logic closely mirrors the cognitive workflow of a radiologist, who typically establishes a baseline of normal anatomical structures before identifying potential abnormalities. By simulating this diagnostic trajectory, the model enhances both the interpretability and the precision of its feature learning process.

To handle the built-in difficulties that come from class imbalance and data noise, the model makes use of purposely arranged regularization methods. The dropout ratios are stepwise adjusted from 0 to 0.1 and at last 0.2, which keeps necessary structure information in the early phases while stronger ratios in the last layers reduce overfitting. In addition, system-level batch normalization is added in all convolution layers to make internal covariate shift steady, thus guaranteeing the network keeps strong resistance to the natural changes and noise that are common in X-ray images.

Ultimately, this structure strikes an optimal equilibrium between feature extraction depth and computational efficiency. Through the synergy of hierarchical learning and rigorous regularization,

DOI: 10.54254/2755-2721/2026.33977

the framework effectively isolates representative pathological features. while maintaining high numerical stability during the training phase.

2.2.2. CNN-transformer hybrid model: design and advantages

Although CNNs have achieved better performance on image recognition tasks, the convolution operations have an innate limitation of not learning global information better, which leads to weak long-term dependencies in CNN-based models [9]. The focus lesions in pneumonia often show themselves as scattered, small, or blurry areas, and their spatial distribution and relation are very important for correct diagnosis. To break this restriction, a CNN-Transformer model is proposed by the present research, which combines the local feature drawing strength of CNNs and the overall context building ability of Transformers. Because a transformer is a deep learning structure that is constructed on a self-attention mechanism, it possesses an advantage in capturing global features and long-range dependence relationships. It can effectively make up for the shortcomings that the traditional CNN model has.

The hybrid structure, which combines the partial feature capturing ability of CNN and the overall situation building of Transformers, hence it solves the complicated problems of pneumonia check. This model obeys a sequence flow pipe: input advance handling, CNN-supported feature drawing out, Transformer-guided global modeling, and full connection classification. This system, which is designed for processing 150×150 single-channel grayscale chest X-ray images, gives a binary classification output that shows either pneumonia or normal states.

In the structural design, the CNN part uses a simplified three-layer construction with increasing channel numbers of 32, 64, and 128. This arrangement makes possible the extraction of hierarchical features, which is supported by Batch Normalization for the stability of training and carefully chosen Dropout rates (0.2 and 0.25) to reduce the problem of overfitting. The following Transformer module is made up of a single encoder layer that has an eight-head self-attention mechanism and a 256-dimensional feed-forward network (FFN). Through bringing in residual connections and Layer Normalization, this module carries out the calculation of attention weights on the feature sequence, enabling the model to distinguish correlations between scattered lesions and the whole anatomical structure of the lungs. After the global feature fusion is completed through the Sequence Average Pooling layer, the compressed data is handled by a dense layer that has 128 neurons and a final Sigmoid activation function. The entire framework is optimized using the AdamW optimizer with a learning rate of 10^{-4} and the Binary Cross-Entropy loss function.

By comparing with traditional pure CNN structure frameworks, this hybrid model provides three main clinical advantages. First, what it realizes is a more excellent combination of fine-grained damage particulars and long-distance space relationships. Second, this method displays promoted stability, keeping its working effect even when there is clinical interference, changes in the placement of patients, or unclear edges of the disease focus. Finally, the model provides improved clinical interpretability by mimicking the diagnostic logic of radiologists, who typically identify local anatomical structures before synthesizing a global pathological assessment.

3. Experiment results and discussion

3.1. Experiment results

To intuitively evaluate classification performance, confusion matrices for both models are presented in Table 3. The data distribution aligns with the inherent class imbalance of medical imaging (1068

DOI: 10.54254/2755-2721/2026.33977

pneumonia vs. 396 normal samples). The Figure 1 shows the confusion matrix of CNN + transformer model. The Figure 2 shows the confusion matrix of CNN model.

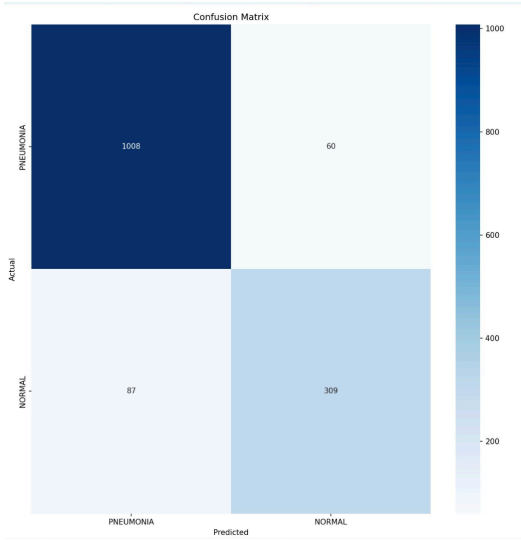

Figure 1. Confusion matrix of the CNN+transformer model (photo/picture credit: original)

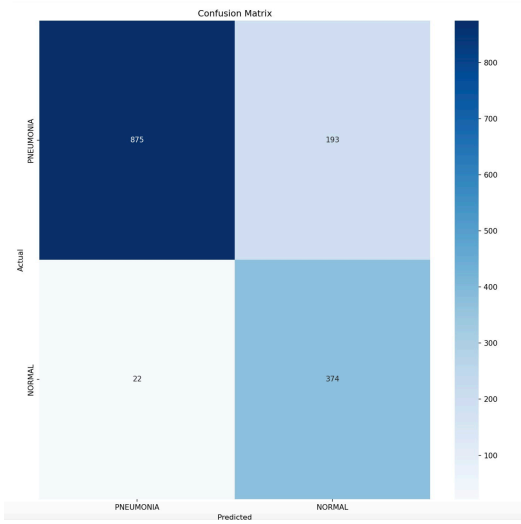

Figure 2. Confusion matrix of the CNN model (photo/picture credit: original)

Table 3. Comparative confusion matrices of the two models

Categories	CNN	CNN+transformer
True Positive	875	1008
False Negative	193	60
False Positive	22	87
True Negative	374	309
Total	1464	1464

In terms of key measurement indices, the CNN+Transformer model obtained a much larger real positive number (1008 vs. 875) and a lower false negative number (60 vs. 193) than the CNN

DOI: 10.54254/2755-2721/2026.33977

model, thus it gives a higher sensitivity (94.38% vs. 81.94 per cent). This shows its stronger capacity for cutting missed diagnosis dangers, which accords with the core clinical need of "prefer false positive over false negative" for the detection of pneumonia. At the same time, the CNN+Transformer model possesses a higher quantity of false positives; hence, it causes a lower specificity value (78.03% vs. 94-point 44 percentage). This trade-off is reasonable and accords with the optimization doctrines of deep learning models, hence it puts first the minimizing of life-endangering missed diagnoses above the cutting down of false alarms.

Accuracy and F1-score give further verification to these results: the CNN+Transformer model has better performance than the CNN model on both correct rate (89.96% vs. 85.38%) and F1-score (93.20 percent versus 89.24%), which can prove that it has a better overall classification effect. This outcome is in accordance with the theoretical superiorities of Transformer structure models in seizing overall features and complicated texture modes, and 9with currently popular conclusions in current medical image analysis studies. The data are displayed within Table 4.

Table 4. Performance comparison of two models

Metric	Recall (Sensitivity)	Specificity	Precision	Accuracy	F1-score
CNN+Transformer	94.38%	78.03%	92.05%	89.96%	93.20%
Pure CNN	81.94%	94.44%	97.55%	85.38%	89.24%

3.2. Discussion and improvement

3.2.1. Model pruning and lightweight optimization

To improve the deployability of the CNN-Transformer hybrid model in clinical environments with limited computing resources, structured model pruning and lightweight optimization are conducted. Redundant parameters in the layers and the fully connected layer are moderately pruned to remove low-contribution weights and unnecessary neurons without obvious performance loss. Meanwhile, less important heads in the multi-head self-attention module are pruned to reduce redundant global calculation. Batch normalization layers are further fused with adjacent convolutional layers to streamline the inference process. These optimizations effectively reduce model size and computational cost, accelerate inference speed, and maintain high accuracy and sensitivity in pneumonia detection, making the model more efficient and suitable for practical clinical application.

3.2.2. Error analysis and practical considerations

The CNN+Transformer model has a smaller quantity of false negative results. This denotes it has a stronger capability for discovering pneumonia-related patterns. It also possesses a greater number of false positive results. This hereby indicates that some normal situations may appear to be pneumonia (for instance, low contrast, overlapped body structure, or imaging-created artifacts).

Such a sensitivity–specificity trade-off is often acceptable in screening/triage settings where minimizing missed cases is prioritized, and the operating point can be further adjusted via decision-threshold selection and probability calibration to meet diverse clinical requirements [10]. In the work that will come after, this research will conduct outside verification on independent groups of samples. This, therefore, can assist in examining the model's robustness in real-world alterations.

DOI: 10.54254/2755-2721/2026.33977

4. Conclusion

This research discusses the shortcomings of conventional single deep learning models for pneumonia examination from chest X-rays and constructs a dependable AI-aided clinical diagnosis instrument. Because pure CNN models have difficulty in capturing long-distance spatial dependence of scattered lesions, a CNN-Transformer hybrid model, which combines the fine-granularity partial feature extraction ability of CNNs and the strong whole context modelling ability of Transformers, is proposed to make up a supplementary and effective disease diagnosis frame.

Experiments were conducted on 5856 chest X-ray images from the Kaggle dataset, these images are divided into training, validation and test set according to the ratio of 60:15:25. Comparing with the pure CNN model, the hybrid model has obtained quite big promotions on the core clinical metrics: sensitivity rises from 81.94% to 94.38%, accuracy rises from 85.38% to 89.96%, and F1-score rises from 89.24% to 93.20%. It greatly decreases the count of missed diagnoses from 193 to 60; this matches the key clinical rule of putting patient safety in first place, hence it keeps an acceptable balance between sensitivity and specificity, therefore it avoids too many false warnings.

To enhance practical performance, image augmentation and model lightweight optimization are applied to improve robustness and enable deployment on resource-limited medical devices. In summary, the proposed model serves as an effective auxiliary tool that eases radiologists' workload, improves diagnostic efficiency, and supports healthcare services in underdeveloped regions. Future research will focus on multi-center dataset expansion, further architectural optimization, and integration into clinical systems to enhance generalization and real-world application value.

References

[1] World Health Organization. (2024). Pneumonia. WHO Health Topics.
[2] Guo, L., Zhou, C., Xu, J., Huang, C., Yu, Y., Lu, G. (2024). Deep learning for chest X-ray diagnosis: Competition between radiologists with or without artificial intelligence assistance. Journal of Imaging Informatics in Medicine, 37(4), 922–934.
[3] Rahman, T., et al. (2020). Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. University of Dhaka.
[4] Kotei, E., & Thirunavukarasu, R. (2024). Tuberculosis detection from chest X-ray image modalities based on Transformer and convolutional neural network. IEEE Access, 12, 97417–97427.
[5] Mathur, M. (2020). Pneumonia detection using CNN (92.6% accuracy). Kaggle.
[6] Widiyanto, W. W., Husen, M. N., Pariyasto, S., & Susanto, E. (2026). Comparative analysis of RGB and grayscale pixel-based similarity methods for lung X-ray image retrieval in clinical decision support systems. Jurnal Teknik Informatika (JUTIF), 6(1), 5854–5864.
[7] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., & Chowdhury, M. E. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Computers in Biology and Medicine, 132, 104319.
[8] Arivoli, A., Golwala, D., & Reddy, R. (n.d.). CoviExpert: COVID-19 detection from chest X-ray using CNN. VIT School.
[9] Fu, B., Zhang, M., He, J., Cao, Y., Guo, Y., & Wang, R. (2022). toHisNet: A hybrid multi-classification model with CNN and Transformer for gastric pathology images. Computer Methods and Programs in Biomedicine, 224, 106924.
[10] Gefter, W. B., Post, B. A., & Hatabu, H. (2022). Commonly missed findings on chest radiographs: Causes and consequences. Chest, 163(3), 650–661.