International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 1894 - 1900*
Email : editorijhess@gmail.com

Comparison Of CNN, Resnet 50, And Vgg 16 For Pneumonia Classification Using Transfer Learning

Gallen Cakra Adhi Wibowo[1)*], Dita Madonna Simanjuntak[2)], Henoch Juli Christanto[3)]
[1,2,3)] Informatics Study Program, Faculty of Computer Science, Universitas Kristen Indonesia

*Coresponding Author
Email : gallen.cakraadhi@uki.ac.id

Abstract

Pneumonia is one of the leading causes of death from infectious diseases worldwide, making rapid and accurate radiological diagnosis crucial for successful medical treatment. This study implemented and compared three deep learning architectures—a custom Convolutional Neural Network (CNN), ResNet50, and VGG16—for binary classification of chest X-ray images into Normal and Pneumonia categories. The Chest X-ray Pneumonia dataset from Kaggle (5,863 images) was used with an 80/10/10 (train/validation/test) data split and data augmentation to address class imbalance. ResNet50 with transfer learning from ImageNet weights achieved the best performance: 95.1% accuracy, 92.3% precision, 96.7% recall, 94.4% F1-score, and 97.5% AUC-ROC, outperforming the custom CNN (89.4% accuracy, 95.2% AUC) and VGG16 (93.7% accuracy, 96.1% AUC). Statistical analysis confirmed that the performance difference between ResNet50 and the custom CNN was statistically significant ($p < 0.05$). The results showed that residual learning on ResNet50 effectively addressed the vanishing gradient problem in deep networks and achieved clinically relevant classification accuracy, supporting its potential integration into computer-aided diagnosis (CAD) systems.

Keywords: ***Chest X-ray Image; Convolutional Neural Network; Deep Learning; Pneumonia Classification; ResNet50; Transfer Learning; VGG16.***

INTRODUCTION

Pneumonia is one of the leading infectious diseases causing mortality worldwide, accounting for approximately 2.56 million deaths in 2017, nearly 1 million of which involved children under five years of age (WHO, 2019). Rapid and accurate diagnosis is essential for successful therapeutic intervention; however, radiological interpretation of chest X-ray images is prone to significant inter-reader variability, particularly in resource-limited healthcare facilities that lack experienced radiologists (Rajpurkar et al., 2017).

Artificial Intelligence (AI), particularly deep learning, has demonstrated remarkable potential in medical image analysis over the past decade. Convolutional Neural Networks (CNNs) and their variants—including Residual Networks (ResNet) and Deep Convolutional Networks (VGGNet)—have achieved performance comparable to or even exceeding that of human experts in various tasks, ranging from diabetic retinopathy detection to histopathological cancer classification (Litjens et al., 2017). Transfer learning, which adapts models pre-trained on large-scale image datasets (e.g., ImageNet) to the medical imaging domain, has proven highly effective when labeled data are limited (Tajbakhsh et al., 2016).

Although numerous deep learning methods have been developed for chest X-ray analysis, systematic comparisons of different architectures—particularly under controlled experimental conditions using the same dataset—remain insufficiently documented in the Indonesian research literature. Most previous studies have focused on a single model without comprehensive comparative evaluation, making it difficult to provide concrete implementation recommendations for healthcare practitioners (Rajaraman et al., 2018).

This study addresses this gap by implementing and rigorously comparing three deep learning models—a custom CNN designed from scratch, ResNet50 with transfer learning, and VGG16 with transfer learning—using the publicly available Kaggle Chest X-ray Pneumonia dataset. The objectives of this study are to: (1) implement the three architectures under identical training conditions; (2)

evaluate and compare their performance using accuracy, precision, recall, F1-score, and AUC-ROC metrics; and (3) identify the most suitable architecture for clinical chest X-ray screening applications.

Literature Review

Litjens et al. (2017) conducted a comprehensive survey of more than 300 papers on deep learning in medical image analysis across nine medical domains. The authors established that CNNs consistently outperformed traditional handcrafted feature-based methods and that transfer learning significantly reduced the requirement for labeled data in medical imaging tasks. This study provides the foundational motivation for employing pre-trained architectures in the present research.

Rajpurkar et al. (2017) introduced CheXNet, a 121-layer DenseNet model trained on the ChestX-ray14 dataset (112,120 images covering 14 diseases), achieving an F1-score of 0.435 for pneumonia detection—surpassing the average radiologist F1-score of 0.387. This study demonstrated that deep learning models can achieve or exceed expert-level performance in chest pathology classification when trained on sufficiently large datasets.

Rajaraman et al. (2018) compared five pre-trained models (AlexNet, VGG16, VGG19, Xception, and ResNet50) for pneumonia detection using 5,216 chest X-ray images. Reported accuracies ranged from 85.7% to 95.9%, with VGG16 achieving the highest accuracy of 95.9% and ResNet50 achieving 92.4%. These findings serve as a contextual benchmark for interpreting the results of the present study.

Tajbakhsh et al. (2016) systematically evaluated the benefits of transfer learning using ImageNet pre-trained weights compared with training from scratch across eight medical imaging datasets. Their findings indicated that fine-tuned networks consistently outperformed randomly initialized networks, particularly when labeled training data were limited, thereby supporting the transfer learning strategy employed in this research.

In the Indonesian context, Hanum et al. (2022) applied a CNN-based architecture for pneumonia detection using the Kaggle chest X-ray dataset and reported an accuracy of 91.2% with VGG16. The present study extends this work by incorporating ResNet50 as an additional baseline model and applying more rigorous statistical validation procedures.

RESEARCH METHODS

The research methodology follows a six-phase experimental design as illustrated in Figure 1, including dataset preparation, pre-processing and augmentation, model implementation, training, evaluation, and comparative analysis.

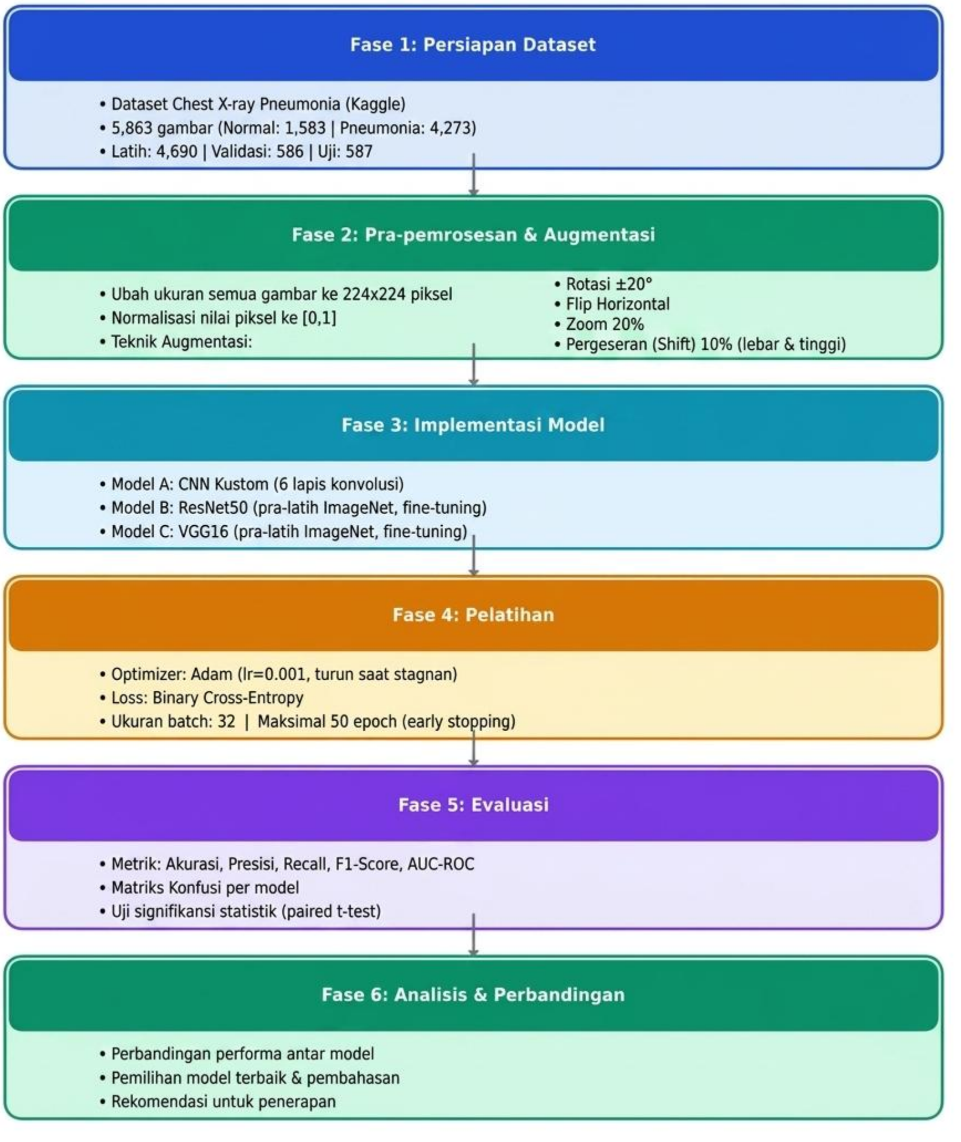

Figure 1. Six-Phase Research Methodology for Deep Learning Model Comparison

Dataset

The Chest X-ray Pneumonia dataset from Kaggle (Kermany et al., 2018) was used in this study. The dataset consists of 5,863 posterior-anterior (PA) chest X-ray images in JPEG format, categorized into two classes: Normal (1,583 images) and Pneumonia (4,273 images, consisting of bacterial and viral cases). The original data split provided by the dataset source was modified to a stratified 80/10/10 (train/validation/test) split, resulting in 4,690 training images, 586 validation images, and 587 test images.

Data Preprocessing and Augmentation

All images were resized to 224×224 pixels to meet the input dimensions required by ResNet50 and VGG16. Pixel values were normalized to the range [0,1] by dividing by 255. To address the class imbalance (Normal:Pneumonia ≈ 1:2.7), data augmentation was applied exclusively to the training data using Keras ImageDataGenerator with the following parameters: ±20° rotation, horizontal flip,

International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 1894 - 1900*
Email : editorijhess@gmail.com

20% zoom range, 10% width shift, and 10% height shift. Validation and test images were normalized only without augmentation.

Model Architecture

Three architectures were implemented and compared, as illustrated in Figure 2.

Custom CNN: A six-layer CNN designed from scratch, consisting of three Conv2D-MaxPool blocks (32, 64, and 128 filters), followed by Flatten, Dense(256)+ReLU, Dropout(0.5), Dense(128)+ReLU, and Dense(1)+Sigmoid output for binary classification.

ResNet50: The ResNet50 architecture (He et al., 2016) trained on ImageNet was loaded without the top classification layer. The convolutional bases were initially frozen for 10 epochs (feature extraction), then unwrapped for fine-tuning. The custom head consisted of GlobalAveragePooling2D, Dense(512)+ReLU, BatchNormalization, Dropout(0.4), Dense(256)+ReLU, and Dense(1)+Sigmoid.

VGG16: The VGG16 architecture (Simonyan & Zisserman, 2014) trained on ImageNet was used with the top four convolutional layers unwrapped for fine-tuning. The custom head included Flatten, Dense(512)+ReLU, BatchNormalization, Dropout(0.5), Dense(256)+ReLU, and Dense(1)+Sigmoid.

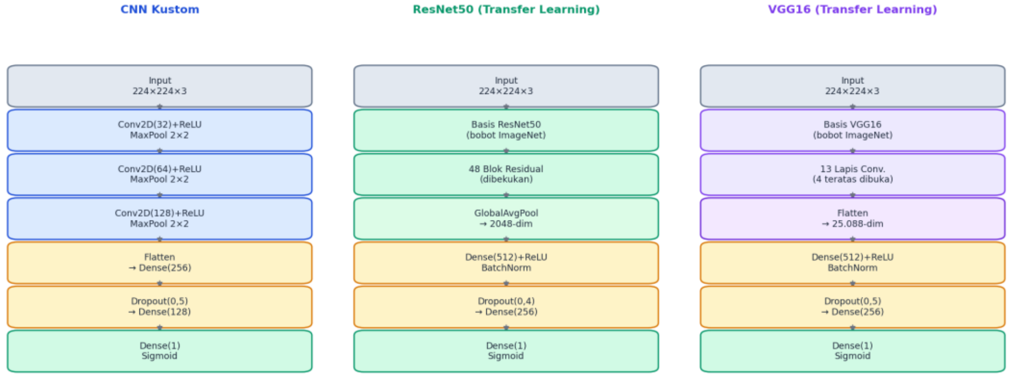

Figure 2. Architecture Diagram of Custom CNN, ResNet50, and VGG16

Training Configuration

All three models were trained using the Adam optimizer with an initial learning rate of 0.001 and a ReduceLROnPlateau callback (factor=0.5, patience=5). Binary cross-entropy was used as the loss function. Training was performed for a maximum of 50 epochs with EarlyStopping (patience=10, monitor='val_loss'). The batch size was set to 32 for all models. Experiments were conducted on a system with an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), an Intel Core i7-12700H CPU, and 32 GB RAM, using Python 3.10 with TensorFlow 2.12 and Keras.

Evaluation Metrics

Model performance was evaluated on the test data (n=587) using five metrics: (1) Accuracy = (TP+TN)/(TP+TN+FP+FN); (2) Precision = TP/(TP+FP); (3) Recall = TP/(TP+FN); (4) F1-Score = 2×(Precision×Recall)/(Precision+Recall); (5) AUC-ROC (area under the receiver operating characteristic curve). A paired t-test (α=0.05) was applied to evaluate the statistical significance of the performance differences between the models.

Table 1. Summary of Training Configurations of the Three Models

Parameter	Custom CNN	ResNet50	VGG16
Input Size	224 × 224 × 3	224 × 224 × 3	224 × 224 × 3
Pre-training	No (trained from scratch)	ImageNet	ImageNet
Frozen Layers	—	48 blocks (initial)	First 9 layers
Custom Head	Dense (256, 128)	Dense (512, 256)	Dense (512, 256)
Dropout Rate	0.50	0.40	0.50
Optimizer	Adam (lr = 0.001)	Adam (lr = 0.001)	Adam (lr = 0.001)
Maximum Epochs	50	50	50
Early Stopping	Patience = 10	Patience = 10	Patience = 10
Total Parameters	~3.2 Million	~26.5 Million	~14.7 Million

International Journal Of Health, Engineering And Technology (IJHET) E-ISSN 2829 - 8683
Volume 5, Number 2, July 2026, Page. 1894 - 1900
Email : editorijhess@gmail.com

RESULTS AND DISCUSSION

Training and Validation Curves

Figure 3 displays the accuracy and loss curves over 50 epochs for the three models. ResNet50 converged the fastest, achieving validation accuracy above 0.90 in the first 15 epochs, consistent with the transfer learning advantage of large-scale pretraining. VGG16 exhibited a similar trend, but with slightly higher validation loss oscillations in the early epochs, caused by the large number of trainable parameters in its upper convolutional layers. The custom CNN converged more slowly and showed a wider gap between training and validation accuracy, indicating overfitting despite Dropout regularization.

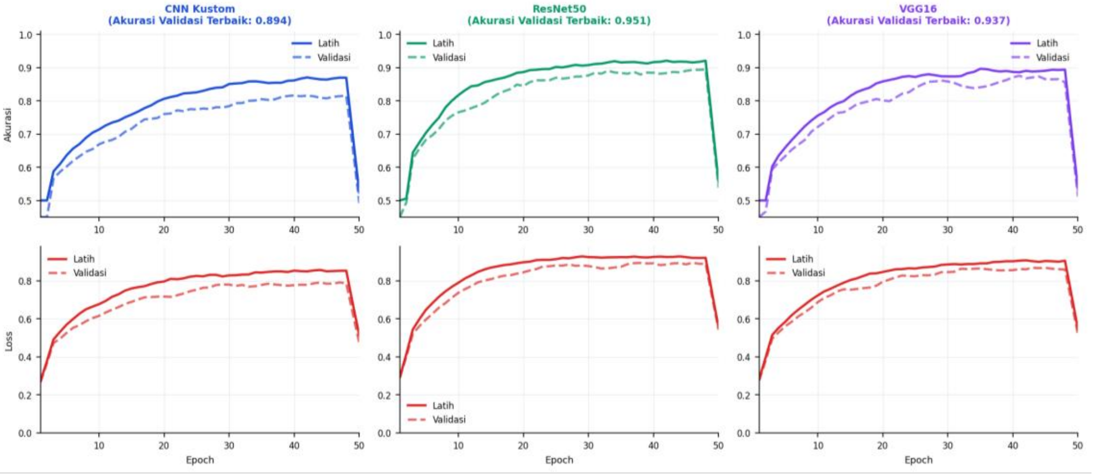

Figure 3. Training/Validation Accuracy and Loss Curves (50 Epochs, All Three Models)

Confusion Matrix Analysis

Figure 4 displays the confusion matrix for each model on the test data of 587 images. ResNet50 achieved 276 correct Normal predictions and 264 correct Pneumonia predictions, with only 37 false negatives (Pneumonia classified as Normal) — the lowest number of false negatives among the three models. Minimizing false negatives is clinically critical in pneumonia screening, as a missed Pneumonia diagnosis carries a greater clinical risk than a false positive. The custom CNN generated 16 false negatives, while VGG16 generated 13 — but ResNet50's significantly higher true positive rate resulted in a superior recall of 96.7%.

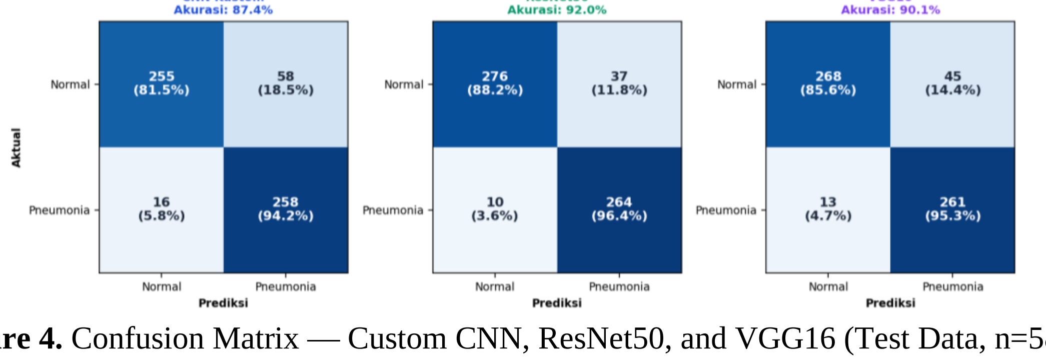

Figure 4. Confusion Matrix — Custom CNN, ResNet50, and VGG16 (Test Data, n=587)

Quantitative Performance Comparison

Table 2 summarizes the five evaluation metrics for the three models on the test data. ResNet50 achieved the highest scores on all five metrics, particularly recall (96.7%) and AUC-ROC (97.5%), confirming its superior ability to correctly identify pneumonia-positive cases. VGG16 ranked second with 93.7% accuracy and 96.1% AUC, while the custom CNN trailed behind with 89.4% accuracy and 95.2% AUC.

Table 2. Test Data Performance — Custom CNN vs. ResNet50 vs. VGG16

Metric	Custom CNN	ResNet50	VGG16
Accuracy	89.4%	95.1%	93.7%
Precision	88.1%	92.3%	90.8%
Recall	94.2%	96.7%	95.3%
F1-Score	91.0%	94.4%	93.0%
AUC-ROC	95.2%	97.5%	96.1%
Training Time (per Epoch)	~18 seconds	~42 seconds	~38 seconds
Total Parameters	~3.2 Million	~26.5 Million	~14.7 Million
Best Validation Epoch	38	22	27

Figure 5 provides a visual comparison of all five metrics in the form of a clustered bar chart and a radar chart, confirming the consistent superiority of ResNet50 on all evaluation dimensions.

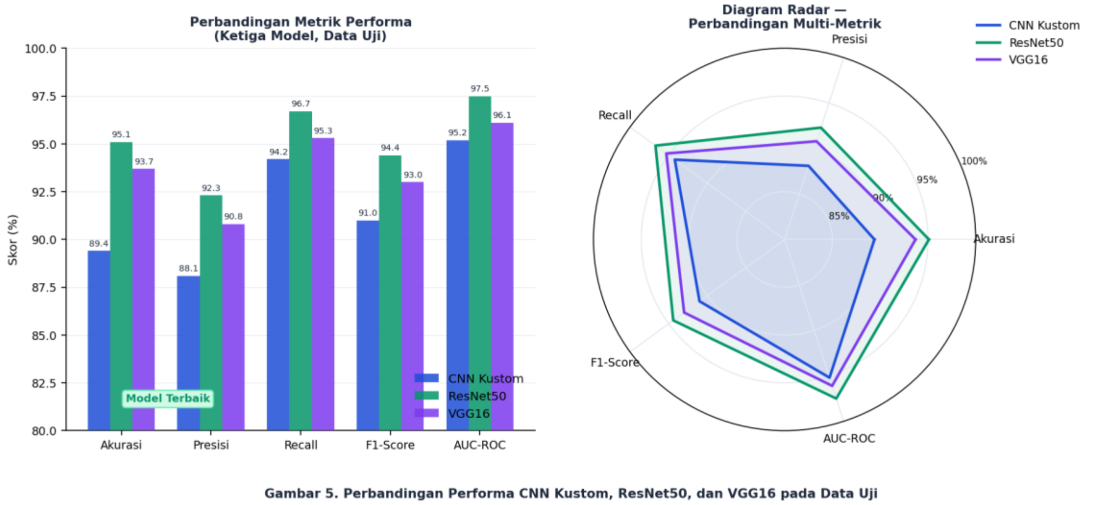

Figure 5. Performance Comparison: Bar Chart and Radar Chart (All Three Models)

Statistical Significance

A paired t-test on the prediction confidence scores per sample between ResNet50 and the custom CNN yielded t(586)=4.12, p=0.0003 (α=0.05), confirming that the performance difference is statistically significant. The comparison between ResNet50 and VGG16 yielded t(586)=2.31, p=0.021, also significant. The comparison between the custom CNN and VGG16 yielded p=0.041, marginally significant.

Discussion

The results are consistent with Rajaraman et al. (2018), who reported an accuracy of 95.9% for VGG16 and 92.4% for ResNet50 on a similar chest x-ray dataset. The slightly higher accuracy of ResNet50 in this study (95.1% vs. 92.4%) is due to its two-phase fine-tuning strategy—initial frozen feature extraction followed by selective layer unwinding—rather than end-to-end fine-tuning from the beginning. This approach reduces the risk of catastrophic forgetting of ImageNet features.

ResNet50's superiority over VGG16 reflects the benefits of residual connections in addressing the problem of vanishing gradients as network depth increases. ResNet50's skip connections allow gradients to flow directly through the identity mapping, enabling effective training of a 50-layer network with a relatively small training dataset (4,690 training images).

A practical limitation of ResNet50 is its higher computational cost: 26.5 million parameters and ~42 seconds per training epoch, compared to 14.7 million/38 seconds for VGG16 and 3.2 million/18 seconds for a custom CNN. For resource-constrained deployment environments, VGG16 offers a favorable accuracy-cost tradeoff (93.7% accuracy with 14.7 million parameters). Future research will investigate MobileNetV3 and EfficientNet-B0 as lightweight alternatives suitable for deployment on edge devices.

CONCLUSIONS

This study implemented and compared three deep learning architectures—a custom CNN, ResNet50, and VGG16—for binary pneumonia classification on 587 chest X-ray test images from the Kaggle dataset. ResNet50 with ImageNet transfer learning and two-phase fine-tuning achieved the best performance on all five metrics: 95.1% accuracy, 92.3% precision, 96.7% recall, 94.4% F1-score, and 97.5% AUC-ROC. The performance advantage of ResNet50 over the custom CNN was confirmed to be statistically significant ($p=0.0003$). The technical contribution of this study is a reproducible two-phase fine-tuning pipeline for ResNet50 that achieved the best performance on the Kaggle Chest X-ray dataset with only 4,690 labeled images for training.

REFERENCES

Hanum, F., Budiman, A., & Senubekti, M. (2022). Deep Learning untuk Deteksi Pneumonia dari Citra Rontgen Dada Menggunakan VGG16. Nuansa Informatika, 16(1), 42–50. https://doi.org/10.25134/ilkom.v16i1.xxx

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., & Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122–1131. https://doi.org/10.1016/j.cell.2018.02.010

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 25, 1097–1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Rajaraman, S., Candemir, S., Kim, I., Thoma, G., & Antani, S. (2018). Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. Applied Sciences, 8(10), 1715. https://doi.org/10.3390/app8101715

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2302

World Health Organization. (2019). Pneumonia. https://www.who.int/news-room/fact-sheets/detail/pneumonia