Deep learning solutions for pneumonia detection: performance comparison of custom and transfer learning models

Yihao Zhong[a], Yanan Liu[b], Erdi Gao[c], Changsong Wei[d], Zhuoyue Wang[e] and Chao Yan[f]*
[a]Courant Institute of Mathematical Sciences, New York University, New York, USA; [b]Research and Development Department, Suzhou Liangchuang Hongzhi Intelligent Technology Co. Ltd., Kunshan, China; [c]Tandon School of Engineering, New York University, New York, USA; [d]Digital Financial Information Technology Co.LTD, Chengdu, China; [e]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA; [f]Department of Electrical and Computer Engineering, Northeastern University
Sunnyvale, USA
* Corresponding author: ycmike97@gmail.com

ABSTRACT

Pneumonia is one of the leading causes of illness and death worldwide. In clinical practice, Chest X-ray imaging is a common method used to diagnose pneumonia. However, traditional pneumonia diagnosis through X-ray analysis requires manual annotation by healthcare professionals which delays diagnosis and treatment. This study aimed to investigate and compare three different deep learning methodologies for classifying pneumonia to detect the disease in patients. These advanced models have the potential to overcome the challenges of reliability and accessibility of diagnostic practices. The methodologies evaluated included a custom convolutional neural network (CNN), a transfer learning approach as well as a fine-tuning strategy based on ResNet152V2. The models were rigorously assessed and compared across various metrics, including testing accuracy, loss, precision, F1 score, and recall. The comparative analysis shows that the fine-tuning strategy outperforms the other methods in terms of operational effectiveness, with the custom CNN being the next most effective, and the transfer learning method ranking last. The study also highlights that false negatives can have more serious consequences than false positives, even without specialized medical knowledge.

Keywords: Pneumonia, deep learning, chest X-ray, ResNet152V2, convolutional neural network

1. INTRODUCTION

Pneumonia, a widespread illness within the medical field, can be provoked by a diverse array of microbial agents, such as bacterial, viral, and fungal pathogens. The term originates from the Greek 'Pneumon' denoting the lungs, thus highlighting its association with pulmonary afflictions. In the medical field, it is recognized as an inflammatory condition that affects lung tissues [1-3]. Besides microbial causes, pneumonia can also arise from the aspiration of food or exposure to harmful chemicals. Alveoli's fluid or pus accumulation due to pathogen's infiltration is the pathogenesis. This leads to disturbances in the alveoli-bloodstream exchange of gases thus subjecting affected individuals to respiratory distress [4-7]. There are various clinical features for pneumonia which includes dyspnea and pyrexia, persistent cough and chest discomfort. Viral types of this disease are highly infectious; they are usually transmitted through close-range droplets and tend to occur in clusters. On time diagnosis of pneumonia may therefore be difficult especially when there is limited access to medical assistance as well as poor transport networks [8-11]. The etiological diagnosis of pediatric pneumonia is further complicated by the low sensitivity of microbiological assays and the nonspecific nature of clinical presentations [12]. Therefore, chest X-ray imaging has become a crucial diagnostic tool that often determines treatment. Conventional diagnostic methods are time-consuming and may vary due to different interpretations by radiologists [13-15]. The interpretation of X-rays is a challenging but important task for radiologists; hence, several computer algorithms have been developed by scientists worldwide. In addition to this, some computer-aided diagnostics (CAD) tools were developed [16-18] to build on these capabilities and improve x-ray interpretation skills for radiologists. Chang's [19-20] work has demonstrated the effectiveness of combining advanced Transformer models with T-SNE dimensionality reduction techniques to build efficient classifiers for fraud detection by effectively capturing complex patterns and dependencies within the data. However, as far as supporting clinicians in making diagnostic decisions during treatment processes is concerned, the accuracy level of these tools is often substandard [21].

International Conference on Automation and Intelligent Technology (ICAIT 2024), edited by R. Usubamatov, S. Feng, X. Mei,
Proc. of SPIE Vol. 13401, 134010E ·
doi: 10.1117/12.3053114

2. METHODOLOGY

2.1 Dataset

We divide dataset into three parts: training, testing, and validation. Each part will have both positive and negative chestX-ray images. The data we used comprises 5,856 pediatric chest X-ray images, sourced from a medical facility in China, encompassing children within the age range of 1 to 5 years. Subsequently, two medical experts conducted a diagnostic assessment of the image samples and assigned corresponding labels to each. To address and validate cases with uncertain diagnostic outcomes, a third expert was engaged to review the evaluation subset. For the purpose of our experiments, we designated 5,232 image samples to the training dataset and set aside 624 images for the testing dataset. 80% was assigned to training and rest 20% for validation. It can be show in table 1.

Table 1. Numbers of Chest X-Ray Images of dataset.

Dataset		
data	Pneumonia	Normal
Training	3106	1079
validation	777	270
testing	396	234
Total no. of pictures	5856	

2.2 Data pre-processing

We established two distinct generators: one dedicated to the rescaling of validation and test datasets, and another for the training dataset, which incorporates additional transformations to augment its size. These generators were then applied to their respective datasets. The parameters employed in the preprocessing of images are outlined. The rescaling operation resizes the images, the width and height shift parameters allow for a 0.1% horizontal and vertical translation, respectively, and the zoom range parameter introduces a random scaling of the image by a factor of 0.1%. Additionally, the zoom range parameter introduces a random scaling of the image to a factor of 0.1%.

2.3 Deep learning model: Custom Convolutional Neural Networks (CNN)

The architecture of the deep learning model designed for classifying pneumonia is illustrated below. This model operates by having each layer utilize the output from its preceding layer as input for further processing. We initiated the model construction by specifying the input parameters to accommodate the dimensions and RGB channels of the incoming sample images. Subsequently, we established the first block: a two-dimensional convolutional layer equipped with 16 filters, each with a 3x3 kernel, accompanied by a batch normalization layer to enhance training efficiency and the model's generalization capabilities. A two-dimensional max pooling layer was then incorporated to condense the feature map and distill the most prominent features, with an additional step of randomly eliminating 20% of the neuron outputs to counteract overfitting.

The role of the first block is to identify foundational image features, such as edges and corners, which are subsequently passed to the second block. The second block mirrors the structure of the first but expands the filter count to 32, funneling its output into the third block. The third block is composed of two convolutional layers, each deploying 64 filters, to delve into more sophisticated feature extraction. This is complemented by batch normalization and max pooling layers, followed by a step that randomly drops out 40% of the neuron outputs. The third block is engineered to capture intricate features necessary for task-specific problem-solving by escalating the complexity of convolutional operations.

Following the initial processing stages, the resultant output undergoes transformation by a Flatten layer, which simplifies the multi-dimensional image data into a streamlined one-dimensional vector. This conversion is crucial for the subsequent layers, which are fully connected and consist of a network of 64 neurons. The ReLU activation function is employed within these layers to add non-linear properties to the model. Additionally, a dropout rate of 50% is implemented on the neuron outputs, serving as a regularization technique to mitigate the risk of overfitting by randomly silencing a subset of neurons during the training phase. These fully connected layers synthesize the extracted features to formulate predictions. The final stage of the model is marked by a single neuron in the output layer, which employs a sigmoid activation function. This function ensures that the output is scaled to a probability value ranging from 0 to 1, which is particularly pertinent for binary classification tasks, such as distinguishing between pneumonia and normal conditions in medical imaging. The resultant output signifies the likelihood of the image indicating a pneumonia-afflicted or normal state.

2.4 Deep learning model: transfer learning

The alternate strategy we employed is the application of the ResNet152V2 architecture within the paradigm of transfer learning. Transfer learning facilitates the migration of knowledge—embodied in the form of parameters—from a seasoned model to a nascent one, thereby augmenting the latter's training endeavors. Given the commonalities that pervade a multitude of image-centric tasks, this technique enables us to equip our new model with the insights garnered from a pre-trained model, enhancing the model's learning efficiency and circumventing the initiation of learning from an entirely naive state.

3. RESULTS

Within the scope of this chapter, an in-depth analysis and comparative assessment of the efficacy of three distinct methodologies for the diagnosis and classification of pneumonia have been undertaken. The training and evaluation phases are uniformly conducted on identical hardware setups. The training process was conducted over 50 epochs, and a checkpoint mechanism was used to save the model parameters that achieved the lowest validation loss. Training is halted and the model reverts to its optimal state if there is no observed reduction in the validation loss over a span of 10 consecutive epochs. Additionally, in the event that the validation loss stagnates without improvement for five epochs, the learning rate is adjusted downward to 20% of its initial value. This strategic reduction aids the model in escaping the confines of a local minimum, thereby facilitating a more extensive exploration of the parameter space for potentially superior solutions.

3.1 Performance analysis

We have deployed a suite of performance metrics to quantitatively assess the efficacy of our deep learning models. Our evaluation encompasses not only the conventional metrics of accuracy and loss across the training, validation, and testing datasets but also employed a confusion matrix to thoroughly evaluate and assess the efficacy of our proposed classification method.

The confusion matrix stands as a pivotal instrument within the machine learning discipline for gauging the proficiency of classification models. It operates as a tabular representation that delineates the discrepancies between the predicted classifications and the true labels of the samples. The matrix's rows correspond to the actual classifications of the samples, while the columns align with the categories as predicted by the model.

The confusion matrix was used to calculate several key performance metrics, such as accuracy, precision, recall (also known as sensitivity), and F1 Score, which were used to evaluate the classification model.

3.2 Custom Convolutional Neural Networks (CNN)

Our model demonstrated a trend of decreasing loss values within the initial 24 epochs, suggesting that it required a relatively shorter duration to attain a good performance of accuracy. The model exhibited a high level of accuracy on the validation set, with a score of 93.9%, and a relatively low loss score of 13.3%. Upon testing with the independent test set, the model achieved an accuracy of 87.8% while incurring a loss of 32.6%, as illustrated in figure 1. The confusion matrix, detailed in figure 2, offers a granular view of the model's classification performance. Our model's precision was calculated at 0.89, with a recall of 0.85 and an F1 score of 0.86, which suggests a well-balanced performance across both precision and recall metrics.

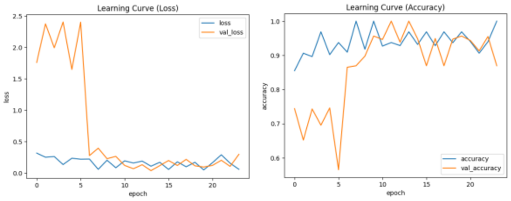

Figure 1. Accuracy and Loss of Custom CNN.

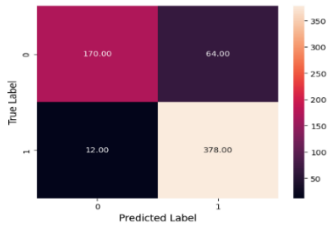

Figure 2. Confusion matrix of Custom CNN.

3.3 Transfer learning

Our model exhibited a reduction in loss values within the span of 16 epochs, indicating a more expedited path to achieving high precision. As a result, the model's efficiency is enhanced, as it requires less computational time to reach a state of high accuracy. In the validation stage, our model delivered a noteworthy accuracy of 95.5%, while incurring a loss of 12.3%. Subsequent testing further demonstrated the model's robustness, with an accuracy of 85.7% and a loss rate of 34%, as shown in Figure 3. The confusion matrix, as presented in Figure 4, offers an in-depth analysis of the model's classification results. The precision of our model was measured at 0.89, the recall at 0.81, and the F1 score at 0.83. This suggests that the model has achieved a robust balance between precision and recall, enhancing its overall predictive performance.

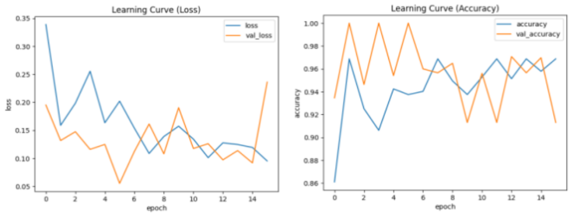

Figure 3. Accuracy and Loss of Transfer learning.

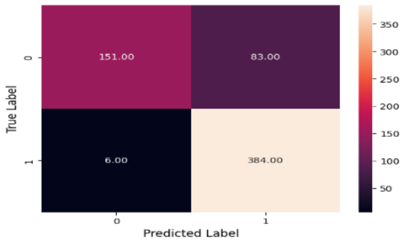

Figure 4. Confusion matrix of Transfer learning.

3.4 Fine tuning

Our model demonstrated a notable decrease in loss values within just 10 epochs, which signifies a more rapid convergence towards high accuracy. Consequently, this efficiency enables the model to reach a state of heightened precision in a shorter amount of time. During the validation phase, our model achieved a high accuracy rate of 95.7%, albeit with a loss score of 19.2%. When the model was subsequently applied to the test set, it maintained a strong accuracy of 90%, albeit with a slightly higher loss of 20%, as illustrated in Figure 5. The confusion matrix, featured in Figure 6, provides an exhaustive analysis of the model's classification efficacy. The model's precision was recorded at 0.91, with a recall of 0.88 and an F1 score of 0.89, which underscores a robust equilibrium between precision and recall in its predictive performance.

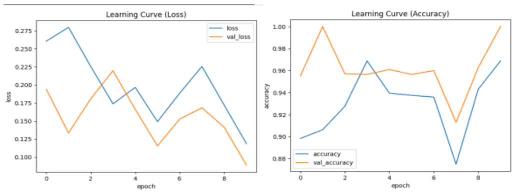

Figure 5. Confusion matrix of Transfer learning.

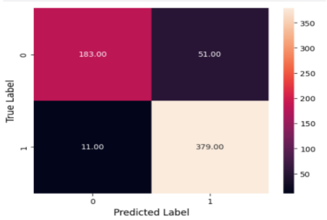

Figure 6. Confusion matrix of Transfer learning.

To sum up, the comparative analysis reveals that the three methodologies under review exhibit a comparable level of performance. When evaluating the models based on parameters such as testing accuracy, testing loss, recall, and the F1 score, the fine tuning approach emerges as the most effective. It is closely followed by the custom CNN, while the transfer learning method ranks as the least effective. In terms of precision, the fine tuning method again stands out as superior, with the custom CNN and transfer learning demonstrating equivalent and consistent performance levels. These findings are further detailed in table 2. Consequently, based on the evaluated criteria, fine-tuning is deemed to be the optimal method when juxtaposed with custom CNN and transfer learning approaches.

Table 2. Comparison of different approach in terms of performance.

Approach	Testing accuracy	Testing loss	Precision	Recall	F1 Score
Custom CNN	0.878	0.326	0.89	0.85	0.86
Transfer learning	0.857	0.34	0.89	0.81	0.83
Fine tuning	0.9	0.28	0.91	0.88	0.89

4. CONCLUSIONS

The study investigated three different deep learning approaches to improve the accuracy of pneumonia diagnosis. Our research encompasses three distinct methodologies: the creation of a custom Convolutional Neural Network (CNN), the utilization of transfer learning techniques, and the execution of fine-tuning strategies. Both the transfer learning and fine-tuning methods are formulated by leveraging the pre-existing knowledge embedded within the ResNet152V2 model. Our study leverages a meticulously selected dataset of 5,232 Chest X-ray images, on which we have applied a series of preprocessing techniques and made strategic adjustments to the network's layers to refine the model's training process. Among the three strategies, fine-tuning has emerged as the most effective, with accuracy rates ranging from 90% to 91%, a recall of 0.88, and an F1 score of 0.89 achieved across 10 epochs of training, which translates to improved precision, recall, and a reduction in computational time. The custom CNN had the second-best performance, while the transfer learning method was the least effective. However, the differences in performance between the three approaches were not significantly large.

We anticipate that through these collaborative technologies, we can attain more refined outcomes across various scenarios. A future objective includes assessing the model's performance across a spectrum of datasets, with the aim of refining its overall accuracy, particularly on datasets pertaining to different pathological conditions. We intend to conduct evaluations on datasets that cover not only pulmonary but also other respiratory ailments. Such endeavors will yield significant insights into the broader utility of our method and its capacity to diagnose a range of related health concerns.

REFERENCES

[1] Jin, Can, et al. "Visual Prompting Upgrades Neural Network Sparsification: A Data-Model Perspective." arXiv preprint arXiv:2312.01397 (2023).

[2] Ding, Z., Li, P., Yang, Q., & Li, S. (2024). Enhance Image-to-Image Generation with LLaVA Prompt and Negative Prompt. arXiv preprint arXiv:2406.01956.

[3] Zhao, H., Lou, Y., Xu, Q., Feng, Z., Wu, Y., Huang, T., … Li, Z. (2024). Optimization Strategies for Self-Supervised Learning in the Use of Unlabeled Data. Journal of Theory and Practice of Engineering Science, 4(05), 30–39. https://doi.org/10.53469/jtpes.2024.04(05).05

[4] Peng, X., Xu, Q., Feng, Z., Zhao, H., Tan, L., Zhou, Y., ... & Zheng, Y. (2024). Automatic News Generation and Fact-Checking System Based on Language Processing. arXiv preprint arXiv:2405.10492.

[5] Lyu, W., Zheng, S., Ma, T., & Chen, C. (2022). A study of the attention abnormality in trojaned berts. arXiv preprint arXiv:2205.08305.

[6] Lyu, W., Zheng, S., Pang, L., Ling, H., & Chen, C. (2023). Attention-Enhancing Backdoor Attacks Against BERT-based Models. arXiv preprint arXiv:2310.14480.

[7] Zhang, X., Wang, Z., Jiang, L., Gao, W., Wang, P., & Liu, K. (2024). TFWT: Tabular Feature Weighting with Transformer. arXiv preprint arXiv:2405.08403.

[8] Liu, R., Xu, X., Shen, Y., Zhu, A., Yu, C., Chen, T., & Zhang, Y. (2024). Enhanced detection classification via clustering svm for various robot collaboration task. arXiv preprint arXiv:2405.03026.

[9] Zhang, J., Wang, X., Ren, W., Jiang, L., Wang, D., & Liu, K. (2024). RATT: A Thought Structure for Coherent and Correct LLM Reasoning. arXiv preprint arXiv:2406.02746.

[10] Ning, Q., Zheng, W., Xu, H. et al. Rapid segmentation and sensitive analysis of CRP with paper-based microfluidic device using machine learning. Anal Bioanal Chem 414, 3959–3970 (2022). https://doi.org/10.1007/s00216-022-04039-x

[11] Zhu, A., Li, K., Wu, T., Zhao, P., & Hong, B. (2024). Cross-Task Multi-Branch Vision Transformer for Facial Expression and Mask Wearing Classification. Journal of Computer Technology and Applied Mathematics, 1(1), 46–53. https://doi.org/10.5281/zenodo.11083875

[12] Shen, Y., Liu, H., Liu, X., Zhou, W., Zhou, C., & Chen, Y. (2024). Localization through particle filter powered neural network estimated monocular camera poses. arXiv preprint arXiv:2404.17685.

[13] Liu, H., Shen, Y., Zhou, W., Zou, Y., Zhou, C., & He, S. (2024). Adaptive speed planning for unmanned vehicle based on deep reinforcement learning. arXiv preprint arXiv:2404.17379.

[14] Mo, Y., Li, S., Dong, Y., Zhu, Z., & Li, Z. (2024). Password complexity prediction based on roberta algorithm. Applied Science and Engineering Journal for Advanced Research, 3(3), 1-5.

[15] Li, Z., Guan, B., Wei, Y., Zhou, Y., Zhang, J., & Xu, J. (2024). Mapping New Realities: Ground Truth Image Creation with Pix2Pix Image-to-Image Translation. arXiv preprint arXiv:2404.19265.

[16] Tao, A., Duan, Y., Wang, H., Wu, Z., Ji, P., Sun, H., ... & Lu, J. (2021). Dynamics-aware Adversarial Attack of 3D Sparse Convolution Network. arXiv preprint arXiv:2112.09428.

[17] P. Huang, S. Park, R. Yan et al., "Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study," Radiology, vol. 286, no. 1, pp. 286–295, 2017.

[18] C. Lyu, H. Yu, Z. Zhao, P. Ji, X. Yang and W. Yang, "Self-Supervised Dense Depth Estimation with Panoramic Image and Sparse Lidar," IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 2023, pp. 6819-6822, doi: 10.1109/IGARSS52108.2023.10283135.

[19] Yu, C., Xu, Y., Cao, J., Zhang, Y., Jin, Y., & Zhu, M. (2024). Credit Card Fraud Detection Using Advanced Transformer Model. arXiv e-prints, arXiv-2406.

[20] Liu, R., Xu, X., Shen, Y., Zhu, A., Yu, C., Chen, T., & Zhang, Y. (2024). Enhanced detection classification via clustering svm for various robot collaboration task. arXiv preprint arXiv:2405.03026.

[21] Zhang, X., Zhang, J., Rekabdar, B., Zhou, Y., Wang, P., & Liu, K. (2024). Dynamic and Adaptive Feature Generation with LLM.. arXiv preprint arXiv: 2406.03505.