Classification of Pneumonia using a convolution neural network model based on Chest X-ray

Gisung Shin[1]

[1]Aiffel, Modulab, 324 Gangnam-daero, Gangnam-gu, Seoul
06252, Republic of Korea

Correspondence: Eunsoo Park, Ph.D
Aiffel, Modulab, 324 Gangnam-daero, Gangnam-gu, Seoul
06252, Republic of Korea
aiffel-cs@modulabs.co.kr

1 Abstract

Background: Pneumonia requires accurate and timely diagnosis for effective treatment. However, conventional methods like chest radiography often suffer from subjectivity and limited efficiency. The performance of data-augmented Convolution Neural Network (CNN) models for pneumonia diagnosis using chest X-rays in pediatric populations remains an area of ongoing investigation and some uncertainty.

Objective: This study aims to develop and evaluate a CNN-based model to classify pediatric chest X-ray images as either pneumonia or normal, and to investigate the impact of data augmentation on model performance.

Methods: A total of 5,863 chest X-ray images from pediatric patients were used, sourced from a public dataset. The dataset was split into training, validation, and test sets. A CNN model was constructed using depthwise separable convolution layers, batch normalization, dropout regularization, and fully connected layers. The model was trained both with and without data augmentation techniques. Performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. Threshold optimization based on F1-score was applied to improve classification balance.

Results: The baseline model (without data augmentation) achieved an accuracy of 87.5%, precision of 87.3%, recall of 93.6%, F1-score of 90.4%, and ROC-AUC of 90.2% after threshold optimization. The data-augmented model outperformed the baseline, achieving an accuracy of 92.95%, precision of 92.2%, recall of 96.9%, F1-score of 94.5 and ROC-AUC of 97.24%. Data augmentation led to a substantial reduction in both false positives and false negatives, enhancing the model's diagnostic reliability.

Conclusion: The implementation of data augmentation significantly improved the CNN model's performance in detecting pneumonia from pediatric chest X-rays. These findings underscore the importance of augmentation in addressing dataset limitations and improving generalization. Further research should focus on external validation across diverse populations and imaging conditions to confirm the model's clinical applicability.

2 Introduction

Pneumonia is a significant global burden of morbidity and mortality, particularly among children and the elderly populations due to their vulnerabilities. It is a well-known form of lung inflammation that primarily affects the alveoli—tiny air sacs essential for gas exchange.[1] In pneumonia, these alveoli become filled with fluid, pus, and cellular debris, leading to airway irritation. In response, the body often triggers a cough reflex to expel these substances—typically caused by bacterial or viral pathogens—in an attempt to ease breathing.[2–4]

Given its high morbidity and mortality, pneumonia requires prompt and accurate diagnosis to enable timely therapeutic interventions and improve clinical outcomes. Accurate classification and early detection are essential in guiding appropriate treatment decisions. However, conventional diagnostic methods, particularly chest radiography, though widely used as a diagnostic cornerstone, present several limitations. Radiographic interpretation can be subjective, highly dependent on the radiologist's expertise, and frequently time-consuming.[5] These limitations may result in delayed diagnosis and treatment initiation, ultimately affecting patient prognosis. Therefore, there is a growing need for more objective, rapid, and standardized diagnostic approaches to enhance clinical decision-making in pneumonia management.[6,7]

Recent developments in deep learning, particularly convolution neural networks (CNNs), have shown great potential in automating and improving the interpretation of chest radiographs. CNN-based models can efficiently learn complex imaging patterns and have been applied to classify various thoracic diseases, including pneumonia, with promising accuracy.[8–10] However, many existing models face challenges related to generalizability, data imbalance, and lack of external validation, limiting their clinical applicability. To address these limitations, this study aims to develop and evaluate a CNN-based classification model for pneumonia using chest X-ray images. Using deep learning techniques, the proposed model seeks to improve diagnostic accuracy and provide a scalable tool to assist clinicians in the early detection and classification of pneumonia.

3 Methods

We used Chest X-Ray Images Datasets for Pneumonia from Kaggle. The image datasets are pediatric patients aged from 1 to 5 years old based on retrospective cohorts and originated from Guangzhou Women and Children's Medical Center in Guangzhou, China. These datasets comprise 5,863 individual images of Chest X-Rays, categorized into two distinct groups: normal and pneumonia. Details described in elsewhere.[10-11]

We built a deep learning model for the pneumonia diagnostic tool using the CNN. To construct this pneumonia diagnostic tool, we approach into three distinct steps: pre-processing, model structure, and classifications.

3.1 pre-processing

The chest X-ray dataset was originally organized into three subsets: training (5,216 images), validation (16 images), and test (624 images) sets. Due to the limited number of validation images, the training and validation datasets were merged, resulting in a combined set of 5,232 images.

This merged dataset was then randomly split into training and validation subsets in an 80:20 ratio, yielding 4,185 images for training and 1,047 images for validation. The test set (624 images) was kept separate and used exclusively for final model evaluation. All images were resized to 180×180 pixels to ensure consistency and efficient processing by the neural network.

To ensure a representative class distribution, the dataset was randomly shuffled prior to the train-validation split. Class labels were inferred from the directory structure: ”normal” indicated healthy cases, while ”pneumonia” denoted infected cases, enabling accurate labeling during pre-processing.

To improve generalization and mitigate overfitting, data augmentation techniques were applied to the training images.[12] These included horizontal flipping, random rotation, zooming, and shifting. In addition, all pixel values were normalized to the range $[0, 1]$.

3.2 Model Structure

The proposed CNN model[13] was designed to classify chest X-ray images as either normal or pneumonia. Input images were RGB chest X-rays resized to a fixed resolution of $180 \times 180 \times 3$. The architecture begins with two standard convolution layers using 16 filters and $3 \times 3 \times$ kernels with ReLU activation, followed by max-pooling. [14-15] Subsequent feature extraction was performed using a series of depthwise separable convolution blocks with increasing filter sizes of 32, 64, 128, and 256.[16] Each block consisted of two SeparableConv2D layers, followed by batch normalization and max pooling to reduce spatial dimensions and improve training stability.[17-18]

Dropout regularization with a rate of 0.2 was applied after the 128- and 256-filter convolution blocks to mitigate overfitting.[19] After flattening, the model included three fully connected (dense) layers with 512, 128, and 64 units, each activated by ReLU and followed by batch normalization and dropout rates of 0.7, 0.5, and 0.3, respectively.[18] The final output layer consisted of a single dense neuron with a sigmoid activation function to perform binary classification.[20]

Indeed, the dataset also have class imbalance with 1,070 images for normal and 3,115 images for pneumonia in the training set.[21-22] Therefore, we computed weights into class to mitigate the potential bias toward the majority class. The formula for computing weight:

$$\text{weight}_c = \frac{1}{\text{count}_c} \times \frac{\text{samples}_t}{2}$$

$$\text{weight}_c = \text{weight for class counts}$$

$$\text{count}_c = \text{number of samples in class counts}$$

$$\text{samples}_t = \text{total number of samples}$$

Both class weights normal (1.96) and pneumonia (0.67) are used to address class imbalance during training. The higher weight for normal compensates for its lower representation in the dataset compared to pneumonia.

3.3 Model classifications

The training process was conducted 25 epochs with a batch size of 16. We monitored validation accuracy and loss to prevent overfitting through early stopping.[23]

The model was compiled using the Adam optimizer and binary cross-entropy loss function, with accuracy as the evaluation metric.[24-25]

The model's performance was evaluated using accuracy, precision, recall, and F1-score on the test dataset. Additionally, a confusion matrix and ROC curve were plotted to analyze the classification effectiveness between pneumonia and normal cases.

3.4 statistical analyses

We analyzed and reported the result of the area under the receiver operating characteristic curve (ROC-AUC), precision, recall, and F1 scores using scikit-learn (version 0.19) via Jupyter.[26-28] These metrics are crucial for assessing the effectiveness of binary classification models.

4 Results

Both baseline and data-augmented CNN models performance evaluations on the unseen test dataset. We used and reported metrics results included loss, accuracy, precision, recall, F1-score, ROC-AUC curve. (Figure 1) Furthermore, confusion matrix and the best threshold optimization were also analyzed for both models.

4.1 Baseline Model Performance

The baseline model (without data augmentation) was evaluated on the test set. Using the default classification threshold of 0.5, the initial evaluation metrics reported a loss of 0.6348, an accuracy of 87.34%, a precision of 86.94%, and a recall of 93.85%. However, the confusion matrix results contradicted the reported precision and recall values. This discrepancy likely arose from incorrect prediction logic that was applied to single probability scores rather than to class probabilities. Consequently, the confusion matrix was deemed unreliable for evaluation at the default threshold.

To address this issue, we analyzed the precision-recall curve and identified an optimal classification threshold of 0.5402 that maximized the F1-score (Figure 2a). After applying this optimized threshold, the model's classification performance improved significantly. The predicted probabilities were better aligned with the true class labels, as reflected in the updated confusion matrix. Evaluation metrics also improved, with an accuracy of 87.5%, a precision of 87.3%, and a recall of 93.6%.

F1-score is the harmonic mean of precision and recall, selecting a threshold that maximizes the F1-score leads to a better balance between these two metrics. This optimized threshold resulted in improved overall classification performance. Indeed, PR and ROC curve reached approximately 89.1% and 90.2%, respectively (Figure 3a and 4a). It implies the model's ability to distinguish between normal and pneumonia cases.

4.2 Data Augmented Model Performance

To enhance the model's generalization capability, a second model was trained using data augmentation techniques, including rotation, shift, zoom, and contrast adjustments. These image transformations aimed to improve robustness against variability in the input data.

The model was first evaluated on the test set using the default classification threshold of 0.5. The initial evaluation metrics demonstrated improved performance compared to the baseline model, with an accuracy of 92.63%, precision of 91.75%, recall of 96.92%, F1-score of 94.26%, and ROC-AUC of 97.24%.

As with the baseline model, we further analyzed the precision-recall curve and identified an optimal threshold of 0.5110 that maximized the F1-score (Figure 2b). Applying this optimized threshold resulted in a slight but not meaningful improvement on performance. However, the data augmentation model consistently outperformed the baseline model across all key metrics. Notably, the augmented model achieved a higher accuracy, precision, recall, F1-score, and a significantly higher on both PR and ROC curve (Figure 3b and 4b).

5 Discussion

We built a convolution neural network (CNN) model to classify chest X-ray images into two categories: normal and pneumonia. The model was implemented using TensorFlow and Keras libraries. Each input image was resized to a fixed dimension of $180 \times 180 \times 3$ to ensure consistency and computational efficiency. The architecture comprises an initial pair of standard convolution layers followed by a series of depthwise separable convolution blocks, each with batch normalization and max pooling. The convolution backbone is followed by fully connected dense layers with dropout regularization. The final output layer consists of a single neuron with a sigmoid activation function to perform binary classification.

The baseline CNN model, trained without data augmentation, was initially evaluated on the test set, yielding a loss of 0.6348, an accuracy of 87.34%, a precision of 86.94%, and a recall of 93.85% with a default threshold of 0.5. To optimize overall classification performance, we analyzed the precision-recall curve and identified an optimal threshold of 0.5402 that maximized the F1-score.[29] This adjustment is crucial as the default threshold of 0.5 is not always ideal for medical diagnostic tasks where balancing false positives and false negatives is critical. After applying this optimized threshold, the baseline model achieved an accuracy of 87.50%, a precision of 87.32%, a recall of 93.59%, and an F1-score of 90.35%. Its ROC-AUC score of 90.22% further indicated a reasonable ability to distinguish between normal and pneumonia cases. However, the confusion matrix revealed a notable number of false positives (53), suggesting challenges in correctly identifying normal cases without misclassify them as pneumonia. This aspect highlights a trade-off that often needs careful consideration in clinical settings.

In contrast, the data augmentation model demonstrated a significant improvement across all evaluation metrics.[30] With an optimized threshold of 0.5110, it achieved a higher accuracy of 92.95%, a precision of 92.20%, a recall of 96.92%, and an F1-score of 94.50%. Critically, its ROC-AUC soared to 97.24%, indicating excellent discriminative power. The confusion matrix for the augmented model showed a substantial reduction in both false positives (32 vs. 53 in baseline) and false negatives (12 vs. 25 in baseline) compared to the baseline, leading to more reliable and balanced predictions (Figure 5). This reduction in false negatives is particularly important in pneumonia diagnosis, as missing a true positive case can have severe clinical consequences.

The significantly enhanced performance of the data augmentation model underscores its effectiveness in improving the robustness and generalization capabilities of the CNN.[31] By introducing diverse variations (random rotation, shift, zoom, and contrast adjustments) to the training images, data augmentation likely support the model learn more invariant features, reducing its susceptibility to specific image characteristics present in the original limited dataset. [32-36] This approach is particularly valuable in medical imaging, where obtaining large, diverse, and annotated datasets can be challenging and resource-intensive. Furthermore, the use of class weighting enable to address data imbalance during training and to promote more balanced learning from both the majority (pneumonia) and minority (normal) classes.

However, it is important to acknowledge the inherent limitations of this study. As the current model was trained and validated exclusively on data sourced from a single medical center (Guangzhou Women and Children's Medical Center), the lack of external validation on independent datasets from diverse patient populations, different geographical regions, and varying equipment/acquisition protocols remains a significant limitation. Such external validation is crucial to fully ascertain the model's clinical applicability and broader generalizability beyond the specific characteristics of the training data.[37-40]

6 Conclusion

In conclusion, data augmentation proved to be a vital technique for improving the diagnostic accuracy of the CNN model for pneumonia detection in pediatric chest X-rays. The augmented model demonstrates a generalization capability for assisting in the diagnosis of pneumonia, offering higher accuracy and more balanced error rates, which are crucial for clinical applications. Future work could explore more advanced augmentation techniques, different CNN architectures, or ensemble methods to further enhance performance.

7 References

1. Hoare Z, Lim WS. Pneumonia: update on diagnosis and management. BMJ. 2006;332(7549):1077-1079. doi:10.1136/bmj.332.7549.1077

2. de Benedictis FM, Kerem E, Chang AB, Colin AA, Zar HJ, Bush A. Complicated pneumonia in children. Lancet. 2020;396(10253):786-798. doi:10.1016/S0140-6736(20)31550-6

3. Jain S, Williams DJ, Arnold SR, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med. 2015;372(9):835-845. doi:10.1056/NEJMoa1405870

4. Kerem E, Bar Ziv Y, Rudenski B, Katz S, Kleid D, Branski D. Bacteremic necrotizing pneumococcal pneumonia in children. Am J Respir Crit Care Med. 1994;149(1):242-244. doi:10.1164/ajrccm.149.1.8111589

5. Langlotz CP. Will Artificial Intelligence Replace Radiologists?. Radiol Artif Intell. 2019;1(3):e190058. Published 2019 May 15. doi:10.1148/ryai.2019190058

6. Mazurowski MA. Do We Expect More from Radiology AI than from

Radiologists?. Radiol Artif Intell. 2021;3(4):e200221. Published 2021 Mar 17. doi:10.1148/ryai.2021200221

7. Waite S, Grigorian A, Alexander RG, et al. Analysis of Perceptual Expertise in Radiology - Current Knowledge and a New Perspective [published correction appears in Front Hum Neurosci. 2019 Aug 13;13:272. doi: 10.3389/fnhum.2019.00272.]. Front Hum Neurosci. 2019;13:213. Published 2019 Jun 25. doi:10.3389/fnhum.2019.00213

8. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. Radiology. 2019;290(2):537-544. doi:10.1148/radiol.2018181422

9. Tang YX, Tang YB, Peng Y, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digit Med. 2020;3(70). doi:10.1038/s41746-020-0273-z

10. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9. doi:10.1016/j.cell.2018.02.010

11. Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

12. Perez, L. and Wang, J. (2017) The Effectiveness of Data Augmentation in Image Classification Using Deep Learning.

13. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS '12). 2012:1097-1105.

14. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9

15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR. 2014;abs/1409.1556.

16. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861. 2017.

17. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proc IEEE Conf Comput Vis Pattern Recognit. 2017;1251-1258.

18. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc 32nd Int Conf Mach Learn. 2015;448-456.

19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929-1958.

20. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006. (or another classic reference for sigmoid activation in classification)

21. He H, Garcia EA. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 2009;21(9):1263-1284. doi:10.1109/TKDE.2008.239

22. King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001;9(2):137-163.

23. Prechelt L. Early Stopping — But When? In: Neural Networks: Tricks of the Trade. Springer; 1998:55-69.

24. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980. 2014.

25. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012. (For binary cross-entropy as a loss function)

26. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861-874.

27. Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. J Mach Learn Technol. 2011;2(1):37-63.

28. Porter P, Brisbane J, Tan J, et al. Diagnostic Errors Are Common in Acute Pediatric Respiratory Disease: A Prospective, Single-Blinded Multicenter Diagnostic Accuracy Study in Australian Emergency Departments. Front Pediatr. 2021;9:736018. Published 2021 Nov 18. doi:10.3389/fped.2021.736018

29. Lipton ZC, Elkan C, Narayanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: Machine Learning and Knowledge Discovery in Databases. Vol 8725. Springer; 2014:225-239.

30. Kumar T, Brennan R, O'Connor NE. Advanced data augmentation approaches: a comprehensive survey and future directions. arXiv. Preprint posted online April 2, 2023. doi:10.48550/arXiv.2304.00445

31. Longjiang E, Zhao B, Liu H, et al. Image-based deep learning in diagnosing the etiology of pneumonia on pediatric chest X-rays. Pediatr Pulmonol. 2021;56(5):1036-1044. doi:10.1002/ppul.25229

32. Xu M, Yoon S, Fuentes A, Park DS. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 2023;137:109347. doi:10.1016/j.patcog.2023.109347.

33. Li X, Wu Y, Tang C, Fu Y, Zhang L. Explicitly learning augmentation invariance for image classification by consistent augmentation. Eng Appl

Artif Intell. 2024;130:107541. doi:10.1016/j.engappai.2023.107541.

34. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. Artif Intell Rev. Published online March 20, 2023. doi:10.1007/s10462-023-10453-z

35. Escobar Díaz Guerrero R, Carvalho L, Bocklitz T, Popp J, Oliveira JL. A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images. J Imaging Inform Med. 2024;37(4):1767-1782. doi:10.1007/s10278-024-01018-9

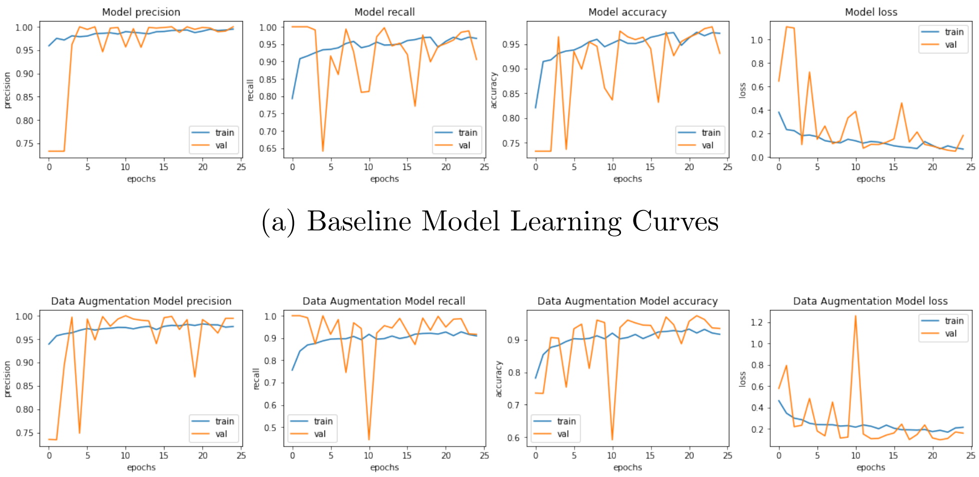
36. Kumar S, Asiamah P, Jolaoso O, Esiowu U. Enhancing Image Classification with Augmentation: Data Augmentation Techniques for Improved Image Classification. Published February 25, 2025. Accessed June 4, 2025.

37. Peled S, Maruvka YE, Freiman M. Multi-Cohort Framework with Cohort-Aware Attention and Adversarial Mutual-Information Minimization for Whole Slide Image Classification. arXiv. Published September 17, 2024. Accessed June 4, 2025.

38. Song YH, Yi JY, Noh Y, et al. On the reliability of deep learning-based classification for Alzheimer's disease: Multi-cohorts, multi-vendors, multi-protocols, and head-to-head validation. Front Neurosci. 2022;16:851871. Published 2022 Sep 7. doi:10.3389/fnins.2022.851871

39. Fathi Kazerooni A, Kraya A, Rathi KS, et al. Multiparametric MRI along with machine learning predicts prognosis and treatment response in pediatric low-grade glioma. Nat Commun. 2025;16(1):340. Published 2025 Jan 2. doi:10.1038/s41467-024-55659-z

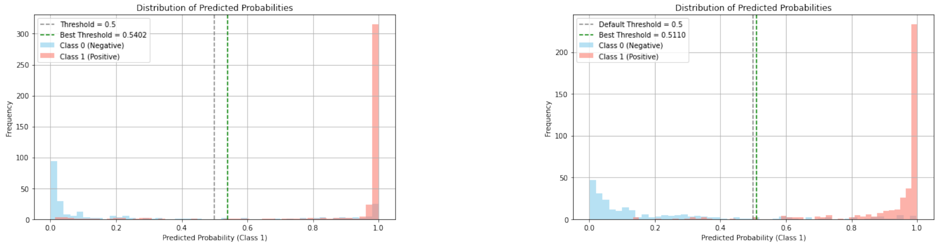
40. Govindarajan ST, Mamourian E, Erus G, et al. Machine learning reveals distinct neuroanatomical signatures of cardiovascular and metabolic diseases in cognitively unimpaired individuals. Nat Commun. 2025;16:2724. doi:10.1038/s41467-025-57867-7



(a) Baseline Model Learning Curves

(b) Data Augmented Model Learning Curves

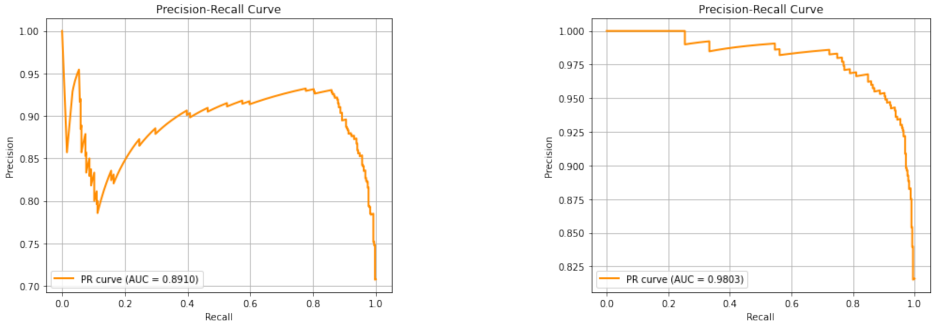
Figure 1: Comparison of Learning Curves Between Baseline and Data Augmented Models



(a) Baseline Model Optimal Threshold

(b) Data Augmented Model Optimal Threshold

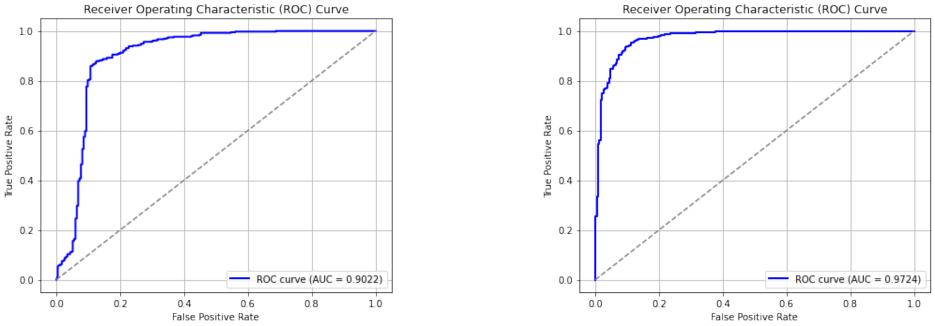
Figure 2: Comparison of Optimal Thresholds Between Baseline and Data Augmented Models



(a) Baseline Model Learning Curves

(b) Data Augmented Model Learning Curves

Figure 3: Comparison of PR-Curves Between Baseline and Data Augmented Models



(a) Baseline Model Result

(b) Data Augmented Model Result

Figure 4: Comparison of ROC-Curves Between Baseline and Data Augmented Model Results

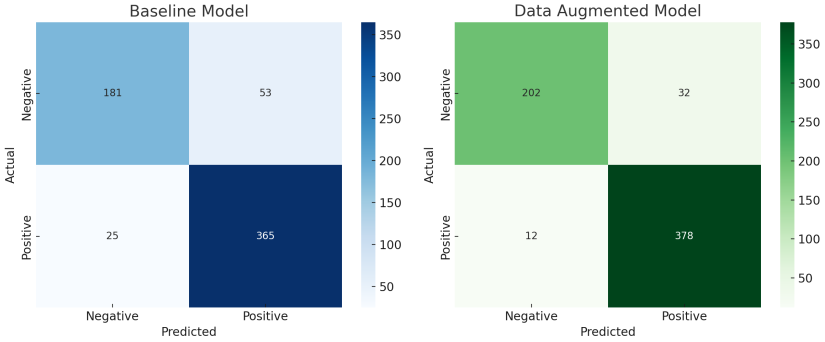


Figure 5: Comparison of Confusion Matrix Between Baseline and Data Augmented Model Results