HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Cross-Domain Transfer Learning for Medical Condition Classification from Infant X-Ray Images

Seyoung Park[1] and Joe Martin[#]

[1]The Webb Schools, USA
[#]Advisor

ABSTRACT

Field of artificial intelligence technology has flourished in recent years which led to a creation of diagnosing or assessing diseases using X-ray images. On the downside, these creations focus mostly on adult X-ray images and can not accurately diagnose infant X-ray images. This issue stems from a combination of factors: limited availability of datasets containing infant X-ray and significant variation in these images due to the rapid development of the infant body. Therefore, there is a high demand to develop comprehensive solutions that address these challenges and provide accurate insights. The proposed representation learning-based framework comprises two stages: auto-encoder-based representation learning and transfer learning for diagnosis. The first stage uses adult X-ray images to train the model for improved representation, generating identical reconstructed images. The second stage utilizes pre-trained models to diagnose diseases and predict infant age, enhancing accuracy by accounting for age-related variations in X-ray shapes. This innovative approach represents the first endeavor in unrestricted pediatric X-ray diagnosis, utilizing self-supervised learning for enhanced accuracy. As a result, the comprehensive and extensive experiment allows the proposed method to outperform in comparison to the existing methods. I expect that my research will contribute to the pediatric field of medicine and serve as the foundation of diverging the utility of artificial intelligence.

Introduction

In recent years, the application of deep learning, particularly convolutional neural networks, has revolutionized the field of medical image analysis. Various studies have demonstrated the efficacy of convolutional neural networks in diagnosis conditions like pneumonia, fractures, and lung diseases from X-ray scans. However, the majority of these models show high accuracy solely for adults while it produces poor results for pediatric X-ray scans. This is because children have developing bodies unlike adults, allowing variety and uniqueness in X-ray scans. As such, to prevent misdiagnosing, there is a necessity to create models that are trained using varieties of pediatric scans. One possible approach to solve this problem is collecting pediatric X-ray scans and creating and training a deep learning model from scratch. However, this procedure will demand a significant amount of time.

As a solution for this problem, I propose a representation learning-based medical condition diagnose framework from infant X-ray images. This proposed method is composed of two different stages: auto-encoder based representation learning and transfer learning aiding medical condition diagnosis. Roughly, the input of the first stage is an adult X-ray image and the output is a reconstructed image of the input. The purpose of this stage is to train the model to create an image that is identical to the input, ultimately, increasing the network performance by achieving better representation and features. Additionally, in the second stage, infant X-ray images are inputted to the pre-trained model and outputs a possible list of diagnosed diseases. In addition to this, the model outputs the predicted age of the infant X-ray as the awareness of the age supports the assessing of the diseases, increasing the accuracy of the model. This is because the shape of the X-ray image can differ

regarding the age of the infant. By utilizing the model trained with adult X-ray images in the first stage, better representation can be found with increased accuracy while enhancing time productivity.

To the best of my knowledge, this is the first attempt in diagnosing pediatric X-ray scans without setting a boundary to certain diseases. As this is the case, there is no previous research and instead the approaches are newly created to better the outcome. Specifically, instead of supervised learning that is commonly used, representation learning or self-supervised learning is used to train the model. This trains the model further as it learns to extract crucial information and output a high-quality reconstructed image, essentially increasing the accuracy of the output.

The organization of this research paper follows: chapter two is the background knowledge that is necessary before progressing through the paper, chapter three works with proposed approach introducing the detail of the method, chapter four contains experiment results, and lastly chapter five is the conclusion that summarizes the research paper with possible future studies.

Background Knowledge

This chapter provides a comprehensive overview of the main concepts and methodologies relevant to the proposed method.

Pediatric X-Ray Scans

There are different types of pediatric X-ray scans, to specify, chest, abdominal, bone, joint, and dental X-rays. Additionally, due to factors including unique body characteristics caused by ongoing developments, the appearance of X-ray scans varies. This means that when radiologists interpret pediatric X-ray scans, they have to understand age-related bone changes. Some of the common disease and conditions that are diagnosed using pediatric X-ray scans includes fractures, bronchiolitis/bronchitis, bronchopneumonia/interstitial pneumonitis, lobar pneumonia, pneumothorax, congenital abnormalities, and foreign body ingestion.

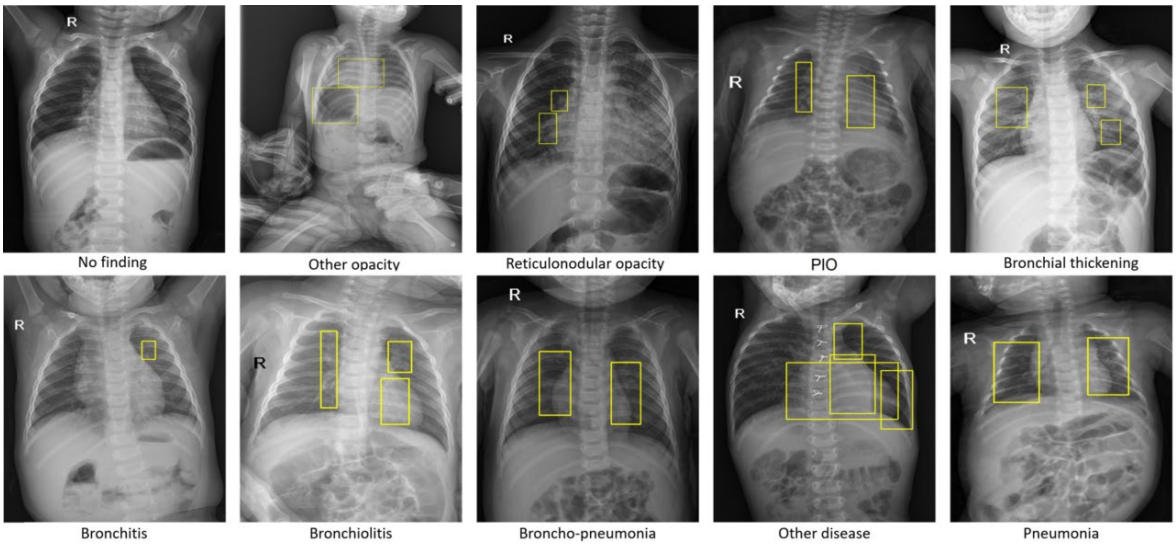


Figure 1. Interpretation of common pediatric diseases diagnosed using X-ray (Pham et al. 2023)

Convolutional Neural Network and Image Classification

Convolutional neural network also known as CNN is a type of neural network used while deep learning. It is widely used while creating models that work with visual data such as image classification. One of the key features of CNN is its convolutional layers. These layers consist of filters that slide over the input image performing element-wise-multiplication to output a feature map. Additionally, some of the most well-known CNN models include: AlexNet, ResNet, and VGGNet.

Image classification is a task in which an input image gets classified into one of the predefined categories, usually performed using convolutional neural networks. Image classification models are trained until they can classify a given image into a specific category with high accuracy.

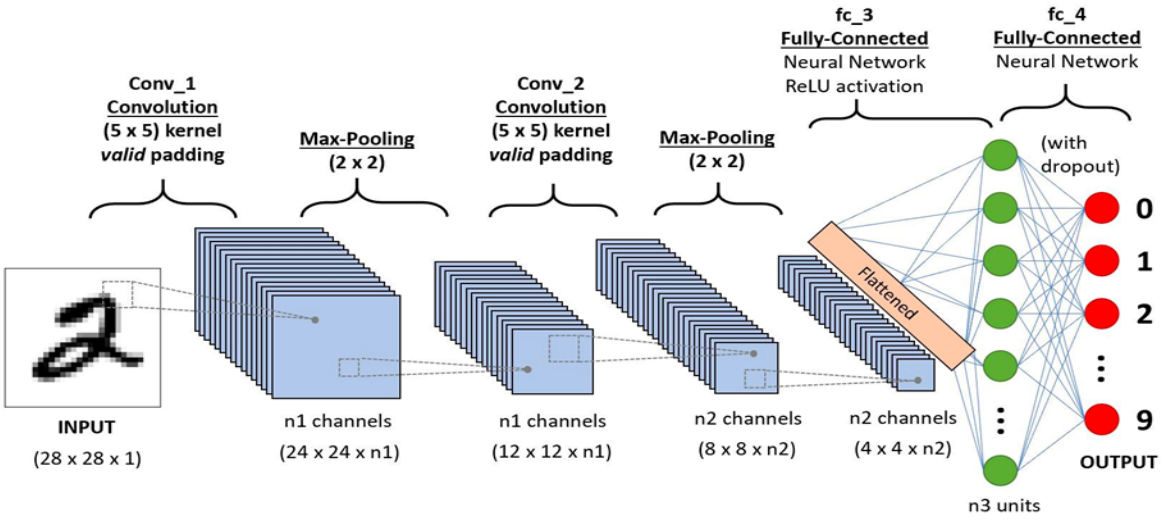
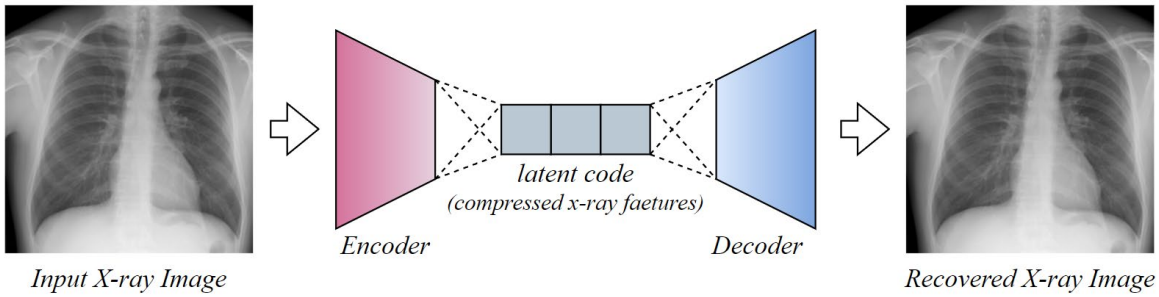
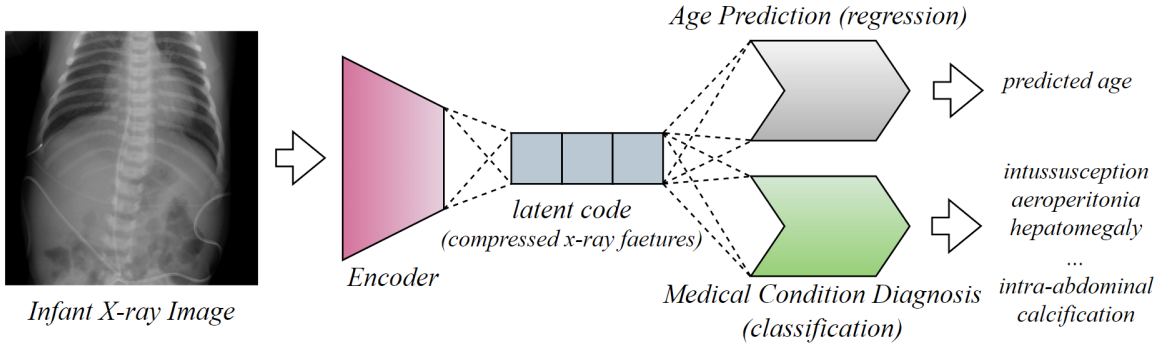


Figure 2. The Basic Framework of Convolutional Neural Network (Bhardwaj 2021)

Proposed Method



(a)



(b)

Figure 3. The Architecture of the Proposed Method. (a): Training the model using Representation Learning and (b): Transfer learning using pre-trained CNN model

The proposed network architecture consists of two main steps, denoted as (a) Training the model using Representation Learning and (b) Transfer learning using a pre-trained CNN model, respectively. Step (a) involves self-supervised learning, where the model is trained using adult X-ray images with the assistance of an auto-encoder. Step (b) utilizes the pre-trained CNN model from (a) to produce outputs, including predictions of an infant X-ray image's age and possible medical condition. Subsequent chapters, 3.1, 3.2, and 3.3, will provide detailed descriptions of X-ray representation learning, transfer learning, and implementation details, respectively.

X-Ray Representation Learning

Fig. (a) in Fig. 3 provides an overview of X-ray representation learning and how the architecture functions. To begin, the input image is denoted as $I \in R^{HW}$. In this notation, the input X-ray, I, belongs to a matrix of HxW, where H and W represent height and width, respectively. The input X-ray image, I, is fed into an encoder, E, represented as $E: I \rightarrow Z$. Here, Z is referred to as the latent code, or simply known as the activation or feature map, which extracts meaningful information from I, essentially compressing the input data.

Next, the latent code is passed into the decoder, D, expressed as $D: Z \rightarrow \hat{I}$. Here, Î represents the reconstructed image, which is the final output of the auto-encoder. The final output from the X-ray representation learning can also be written as $\hat{I} \in X^{HW}$.

Repeatedly performing this self-supervised learning creates a pre-trained CNN activation or feature map with enhanced information, which will be advantageous when performing transfer learning later on.

Transfer Learning

Fig. (b) in Fig. 3 introduces transfer learning using the pre-trained CNN model from Fig. (a). Firstly, instead of the adult X-ray image used in the representation learning, an infant X-ray image serves as the input, denoted as I. Similarly, it is passed through an encoder known as the pre-trained CNN model, E, represented as $E: I \rightarrow Z$. Z is the latent code that provides a better representation utilizing representation learning. Subsequently, the information enters the neural network called AgePredictionNet, abbreviated as APN, to output the predicted

age and potential diagnosed medical conditions (MCD). This is represented as $APN: Z \rightarrow Age$ / $MCD: Z \rightarrow P$. Here, *Age* and *P* are the final outcomes.

Transfer learning, as mentioned in this architecture, reduces machine training time while enhancing performance. This is achieved by leveraging the network created during the representation learning. In summary, this two-step network architecture can minimize the loss function and provide closer-to-accurate outputs.

Implementation Details

The fourth section of this research paper will present the application of this method and provide an illustration of the experiments conducted. It will showcase the accuracy of the two-step learning architecture, featuring both representation learning and transfer learning. Below this are the loss functions that were used during the analysis that calculates that loss of representation learning, difference between the predicted and the real diagnosis, the Age Prediction Regression, and the total loss value, respectively.

Equation 1: L1 Loss Function

$$L_1 = \frac{1}{HW} \sum_x^W \sum_y^H |I_{(x,y)} - \hat{I}_{(x,y)}|$$

In the equation 1, *H* and *W* represent the height and width of the inputted image and *I(x,y)* denotes the pixel intensity of a certain location in the picture. While *I(x,y)* represents the predicted value of a certain location. The equation calculates for the absolute value of the difference of all the pixel values showing how close the predicted values are in comparison to the real values.

Equation 2: Cross-Entropy Loss Function

$$L_{CE} = - log_e P$$

The equation 2 quantifies the difference between the predicted and the real diagnosis. ***P*** represents the predicted probability of the input diagnosis and the -log function converts the logarithm of the probability into a positive value. When the *P* is closer to 1, meaning the model is confident that the correct diagnosis is being predicted), the negative natural logarithm of *P* becomes close to 0, resulting in a low loss. Conversely, when *P* is close to 0 (meaning the model is confident that incorrect diagnoses are being predicted), the negative natural logarithm of *P* becomes a large positive value, resulting in a high loss.

Equation 3: Loss Function for Age Prediction Regression

$$L_{MSE} = \frac{1}{N} \sum_1^N |age_i - \widehat{age}_i|^2$$

The mean square error function, derived from the L_1 loss function, calculates for the absolute value of the difference between the real age of the infant, age_i, and the predicted age of the infant, $\widehat{age}_i$.

Equation 4: Final Loss Value

$$L_{total} = L_{CE} + \alpha \cdot L_{MSE}$$

In equation 4, the final loss value, including both the loss of the representation learning and the Age Prediction Regression is calculated. Alpha (α) is a decimal value; in this case: 0.8, that decreases the importance of the loss value of the Age Prediction.

Experimental Results

Dataset

The size and shape of the abdominal cavity in infants vary with months, exhibiting significant differences from the abdominal cavity of adults. Therefore, the patterns of abdominal pain commonly experienced in adults differ from those of children. Consequently, applying deep learning models trained on adult abdominal X-rays to children presents limitations. As for the situation, while training the model using representation learning, 241,562 adult X-ray datasets with no labels and 70,353 infant X-ray datasets categorized into 11 different labels were both used. This allowed the model to compress X-rays of adults while familiarizing with infant X-ray images as well. While training the model using transfer learning in stage 2, solely infant datasets were used. Since the objective of this study is to diagnose infant X-rays, by performing such transfer learning, models can demonstrate the result with better representation. Specifically about the infant datasets that were used, these data were collected from AI-hub (AI Hub 2023). The 11 different labels include: normal, aeroperitoneum, mass, intussusception, air-liquid shadow, foreign bodies, constipation, calcification, splenomegaly, hepatomegaly, and situs inversus totalis.

Evaluation Metrics

Confusion Matrix

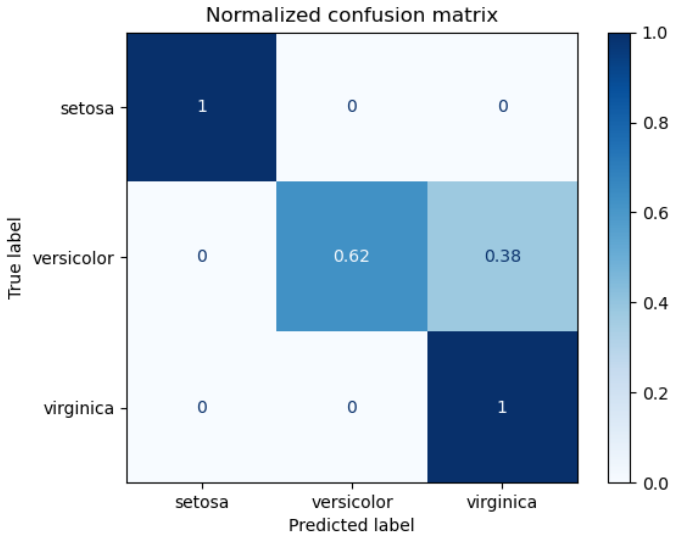


Figure 4. Example of visualization of confusion matrix

Figure 4 is a visualization of the confusion matrix. Confusion matrix is a common tool that is used to evaluate the performance of the machine learning model. It summarizes the experimental results by breaking down the result into four different key components: true positives (TP), true negative (TN), false positive (FP), and false negative (FN). In these circumstances, TP represents instances where AI correctly predicted positive results,

TN represents instances where AI correctly predicted negative results, FP represents instances where AI incorrectly predicted the result as positive, and lastly, FN represents instances where AI incorrectly predicted the positive result as negative.

Equation 5: Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

This equation calculates the accuracy of the model using the four major components of the confusion matrix. The denominator denotes all the samples collected regarding the AI model's performance and the numerator denotes the combined values of TP and TN showing the accuracy of the AI model.

Equation 6: Recall

$$\frac{TP}{TP + FN}$$

This equation calculates the positive sample that AI correctly predicted out of all the positive samples. The denominator represents all the positive samples including the ones that the AI model incorrectly predicted while the numerator is the TP values.

Equation 7: Precision

$$\frac{TP}{TP + FP}$$

This equation calculates for the precision of the AI model. The denominator is all the data regarding TP and FP, showcasing the positive real values while the numerator is the TP, the values that the AI model predicted correctly as a positive value

Equation 8: F1-score

$$\frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

F1-score is the harmonic mean that calculates for a single number that combines both precision and recall for a balanced measure of the model's performance. Thus, it evaluates the overall accuracy of the classification model using recall and precision.

ROC Curve

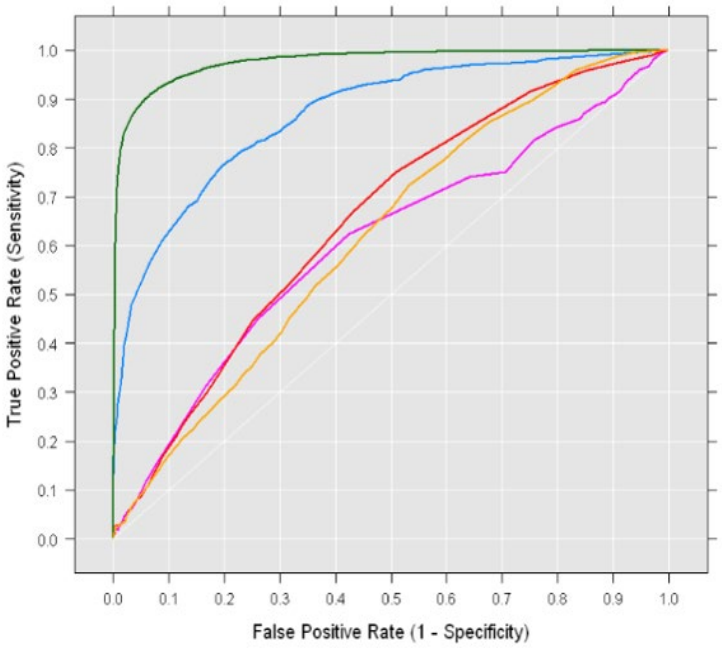


Figure 5. Example of ROC curve

A Receiver Operator Characteristic (ROC) curve is a graphical representation of binary classifiers often using true positive rate (TPR) as a y-axis and false positive rate (FPR) as a x-axis. Previous to the creation of the graph, threshold value is selected for the improvement of the classification model in which it can evaluate the TP, TN, FP, FN. The threshold value is diversified so that the collection of different TP, TN, FP, and FN values are possible. This data is then taken to the consideration, compared and analyzed, leading to the ROC curve. As the picture demonstrates, as the curve gets closer to the left top corner, it indicates a successful model. Ideally, the graph should go along the line of the y-axis and the top like a rectangular box, however if it's a random model, it will be graphed as a straight line.

Performance Comparison

Table 1. Two groups broken down with age ranges and the difference.

	Accuracy	**Recall**	**Precision**	**F1-Score**
VGG19 (Simonyan et al. 2014)	0.8050 (±0.0014)	0.8059 (±0.0013)	0.8041 (±0.0009)	0.8050 (±0.0012)
MobileNetV2 (Sandler et al. 2018)	0.8067 (±0.0011)	0.8060 (±0.0012)	0.8038 (±0.0008)	0.8049 (±0.0010)
ShuffleNet (Zhang et al. 2018)	0.8084 (±0.0012)	0.8101 (±0.0010)	0.8104 (±0.0008)	0.8102 (±0.0009)

Xception (Fran et al. 2017)	0.8080 (±0.0014)	0.8111 (±0.0011)	0.8110 (±0.0015)	0.8110 (±0.0012)
HRNet-w32 (Wang et al. 2020)	0.8160 (±0.0008)	0.8191 (±0.0007)	0.8185 (±0.0011)	0.8188 (±0.0013)
ResNext-50 (Xie et al. 2017)	0.8190 (±0.0007)	0.8221 (±0.0008)	0.8233 (±0.0013)	0.8227 (±0.0009)
Densenet-121 (Huang et al. 2017)	0.8210 (±0.0011)	0.8231 (±0.0013)	0.8212 (±0.0008)	0.8221 (±0.0011)
Resnet-50 (He et al. 2016)	**0.8249 (±0.0013)**	**0.8280 (±0.0008)**	**0.8251 (±0.0011)**	**0.8250 (±0.0012)**

As mentioned, this experiment on using transfer learning for medical condition classification, specifically diagnosing infant X-rays is conducted to help diagnose infants both accurately and quickly while proving the benefits of using transfer learning in convolutional neural network (CNN) models. In order to examine the proposed method, I experimented with a total of eight different networks. By doing so, a comparison of network performance is possible, allowing the exhibition of the most reliable network. Among these networks, VGG19 and MobileNetV2 performed relatively poorly compared to other networks while ShuffleNet and Xception performed slightly better than those two. This is the consequence of their shallow convolutional layers, which limit the ability to capture and learn complex patterns and are inadequate for dealing with complex data such as diverse infant X-ray images. Contrastingly, as shown by the graph, ResNext-50, Densenet-121, and Resnet-50 exhibited slightly improved performance. These networks had deeper convolutional layers which allowed capturing and storing complex, meaningful representation data. Ultimately, Resnet-50 with its 50 convolutional layers performed the best among the eight.

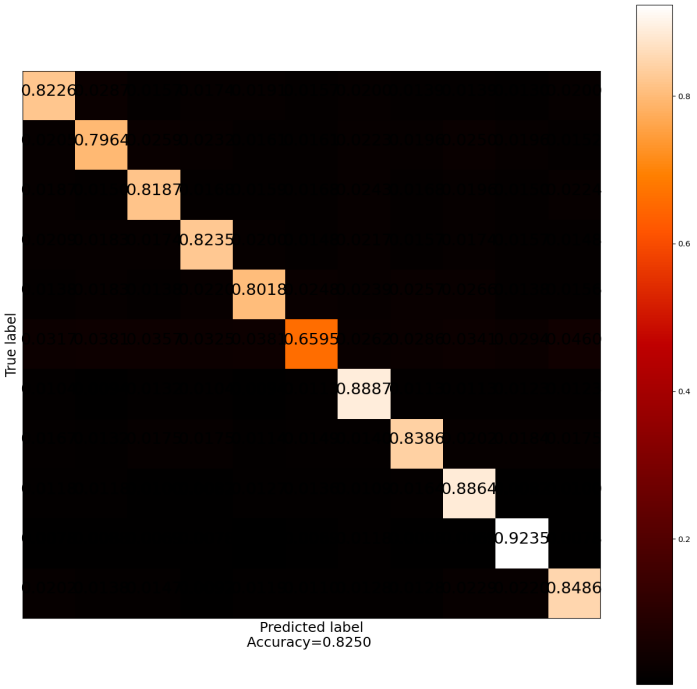


Figure 6. Confusion matrix of the proposed method

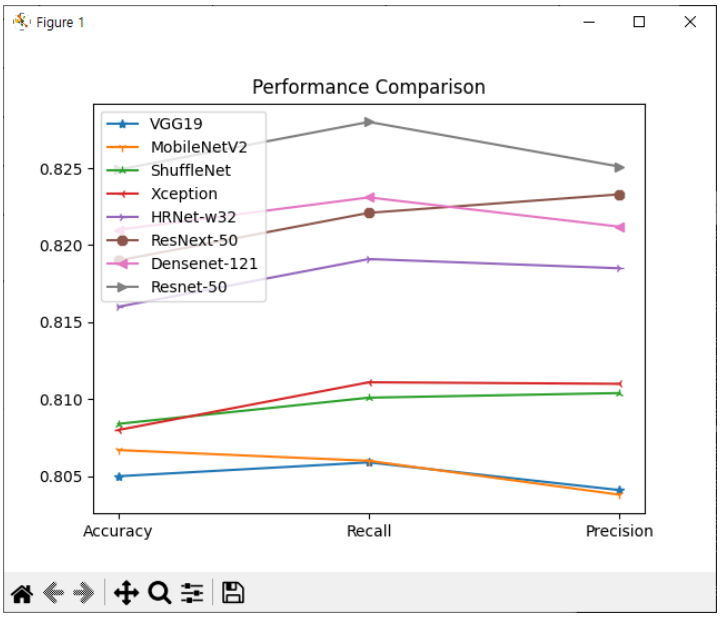


Figure 7. Performance comparison (line graph)

The confusion matrix additionally represents the accuracy of the networks in a matrix format. The y-axis represents the true label while the x-axis represents the predicted label and the gradation scale on the right showcases the chart demonstrating accuracy, from lightest to the darkest, most accurate to least accurate, respectively. Diagonal components in the matrix are the ratio of correct network predictions. As shown, most of

the diagonal components consist of lighter hues conveying the accuracy of the networks in predicting age and diagnosing infant X-rays.

Ablation Study

Table 2. Ablation study result

	Accuracy (ablation model-1)	**Accuracy (ablation model-2)**	**Accuracy (Full model)**
VGG19 (Simonyan et al. 2014)	0.7772 (±0.0009)	0.7927 (±0.0008)	0.8050 (±0.0014)
MobileNetV2 (Sandler et al. 2018)	0.7847 (±0.0011)	0.7957 (±0.0009)	0.8067 (±0.0011)
ShuffleNet (Zhang et al. 2018)	0.7732 (±0.0009)	0.7910 (±0.0008)	0.8084 (±0.0012)
Xception (Fran et al. 2017)	0.7669 (±0.0012)	0.7884 (±0.0011)	0.8080 (±0.0014)
HRNet-w32 (Wang et al. 2020)	0.7765 (±0.0013)	0.7974 (±0.0010)	0.8160 (±0.0008)
ResNext-50 (Xie et al. 2017)	0.7728 (±0.0009)	0.7965 (±0.0007)	0.8190 (±0.0007)
Densenet-121 (Huang et al. 2017)	0.7786 (±0.0010)	0.7954 (±0.0012)	0.8210 (±0.0011)
Resnet-50 (He et al. 2016)	**0.7797 (±0.0011)**	**0.8004 (±0.0009)**	**0.8249 (±0.0013)**

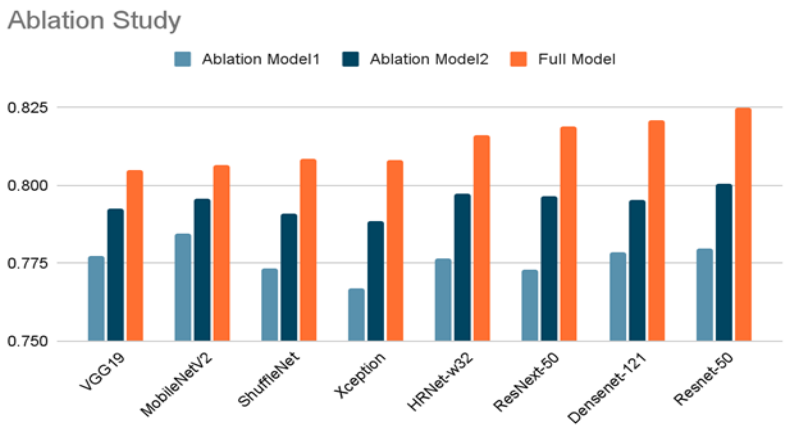


Figure 8. Ablation study result (bar graph)

The purpose of the ablation study is to examine the importance of certain elements utilized in the proposed method, specifically the effects of applying representation learning and inserting AgepredictionNet into this specific model. In order to visualize the effect, two different ablation models were proposed as shown by the graph. The bar graphs in the figure explicate the difference in the accuracy of each network in regards to utilizing two different ablation models compared to the full model. Clearly, the model without the representation learning performed severely worse than the full model while the model without the AgepredictionNet performed relatively worse. This can be attributed to the fact that representation learning helps with extracting specific and complex data well and, considering the amount of infant dataset, transfer learning helps with increasing the performance level while handling limited datasets. This data reveals the importance of utilizing elements that are used in machine learning and the potential of this new model compared with other models.

Conclusion

In this research study, I have proposed a cross-domain transfer learning approach for the classification of medical conditions, specifically in the context of diagnosing infant X-ray images. The proposed method utilizes representation learning and convolutional neural networks to extract crucial data, which is then used in transfer learning to enhance the performance of the machine with a limited dataset of infant images. Through extensive experiments in machine learning, I have demonstrated that the self-supervised representation learning approach not only outperforms supervised representation learning but also enhances the effectiveness of transfer learning in extracting new features for diagnosing foreign images.

Furthermore, the performance study involving different CNN architectures highlights the importance of utilizing models with deep convolutional layers when dealing with complex data, leading to improved accuracy. Moving forward, I aim to further develop the proposed method to achieve higher accuracy in classifying infant X-ray images with other diseases in specific, ultimately bridging the gap between the fields of machine learning and healthcare for a better future in terms of quality healthcare.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

AI Hub (2023, May 1). "*Pediatric abdominal x-ray image data*": AI Hub https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71432

Bhardwaj, Shivam. (2021, Jun 8). *"Convolutional Neural Networks : Understand the Basics"*: Analytics Vidhya https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-understand-the-basics/

Fran, C. (2017). Deep learning with depth wise separable convolutions. In IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.48550/arXiv.1610.02357

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.48550/arXiv.1512.03385

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708). https://doi.org/10.48550/arXiv.1608.06993

Pham, H. H., Nguyen, N. H., Tran, T. T., Nguyen, T. N., & Nguyen, H. Q. (2023). PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. Scientific Data, 10(1), 240.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520). https://doi.org/10.48550/arXiv.1801.04381

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364. https://doi.org/10.48550/arXiv.1908.07919

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500). https://doi.org/10.48550/arXiv.1611.05431

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856).
https://doi.org/10.48550/arXiv.1707.01083w