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

Deep Learning for Neurodevelopmental Disorder Diagnosis: Leveraging Retinal Images with Self-Supervised Learning

Beom Kim[1], Hwan Kim[1] and Sojung Min[#]

[1]Korean Minjok Leadership Academy, Republic of Korea
[#]Advisor

ABSTRACT

Neurodevelopmental disorders-like autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) have posed substantial challenges to the cognitive, social, and emotional development of children. Early and precise diagnosis of ASD and ADHD holds significant importance in facilitating timely interventions and support, ultimately enhancing outcomes for individuals affected by these conditions. Conventional diagnostic approaches for these disorders have depended on behavioral assessments, clinical observations, and structured interviews, which are labor-intensive and susceptible to subjective judgments made by healthcare professionals. In this research paper, we proposed a deep learning-based system for diagnosing neurodevelopmental disorders. The proposed system utilizes retinal images as input and generates disorder predictions. We introduce a novel self-supervised learning approach to attain superior accuracy outcomes. We demonstrated that the proposed system achieves state-of-the-art accuracy on large-scale retinal image datasets with extensive experimental results.

Introduction

Neurodevelopmental Disorder

Autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD), the common neurodevelopmental disorders present challenges to individuals, their families, and society. These disorders often manifest early in life, affecting cognitive, social, and emotional development. Timely diagnosis and intervention are crucial to improving the outcomes and quality of life for affected individuals. Traditionally, the diagnosis of these neurodevelopmental disorders has mainly relied on behavioral assessments such as clinical observations, and structured interviews conducted by experienced healthcare professionals. While these methods are valuable and provide accurate results, they are labor-intensive, time-consuming, and subject to variations in judgments. As a result, there is a growing need for more efficient and objective diagnostic approaches.

Retinal Image

A retinal image is a digital image of the back of human eyes. It shows the optic nerves which send information to the brain. Examining retinal images can provide various information about patients' overall health conditions (Patton et al. 2006). While retinal imaging is a non-invasive inspection, it provides the capability to visualize retinal blood vessels. Due to this unique feature, there haven’t been research studies attempting to diagnose neuro-related diseases such as Parkinson's disease.

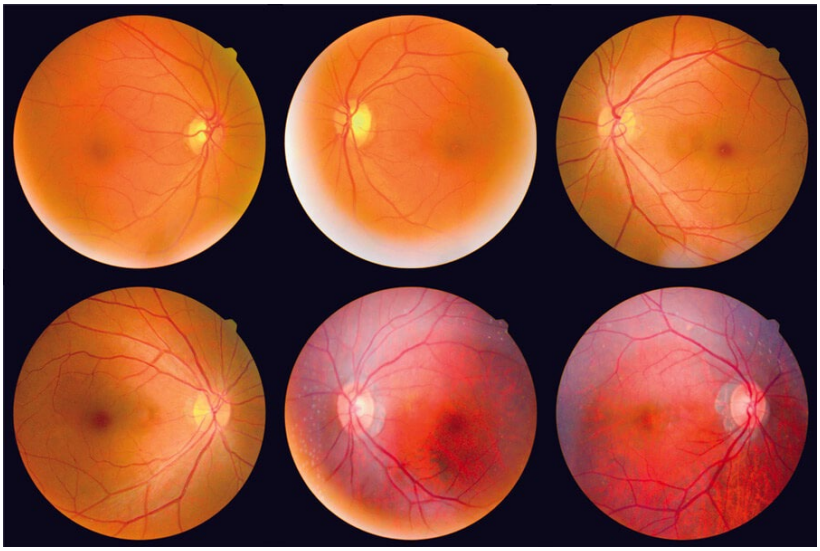

Figure 1. Examples of retinal images

Previous study introduced a machine learning-based approach for diagnosing Parkinson's disease using retinal images (Zhou et al. 2023). Their research involved the analysis of visual characteristics in retinal blood vessels and uncovered a potential correlation between retinal images and Parkinson's disease, a neuro-related condition.

Inspired by their work, we propose a deep learning-based system designed for the diagnosis of neurodevelopmental disorders. The proposed system takes retinal images as input and produces disorder predictions as output. We also introduce a novel self-supervised learning approach to enhance accuracy. Furthermore, we implement an age and gender-guided approach to improve performance. A comprehensive description of the proposed system is provided in Chapter 3.

Proposed Neurodevelopmental Disorder Diagnosis System

In this chapter, we provide an overview of the architecture of the proposed neurodevelopmental disorder diagnosis system. The proposed approach leverages representation learning, which aims to create meaningful features from retinal images. During the representation learning phase, the trained network learns to extract significant and relevant visual patterns from the retinal images, specifically those related to blood vessels. Subsequently, we employ transfer learning or fine-tuning techniques to implement the neurodevelopmental disorder diagnosis system. Chapters 2.1 and 2.2 will provide detailed explanations of each component of the system.

Retinal Representation Learning

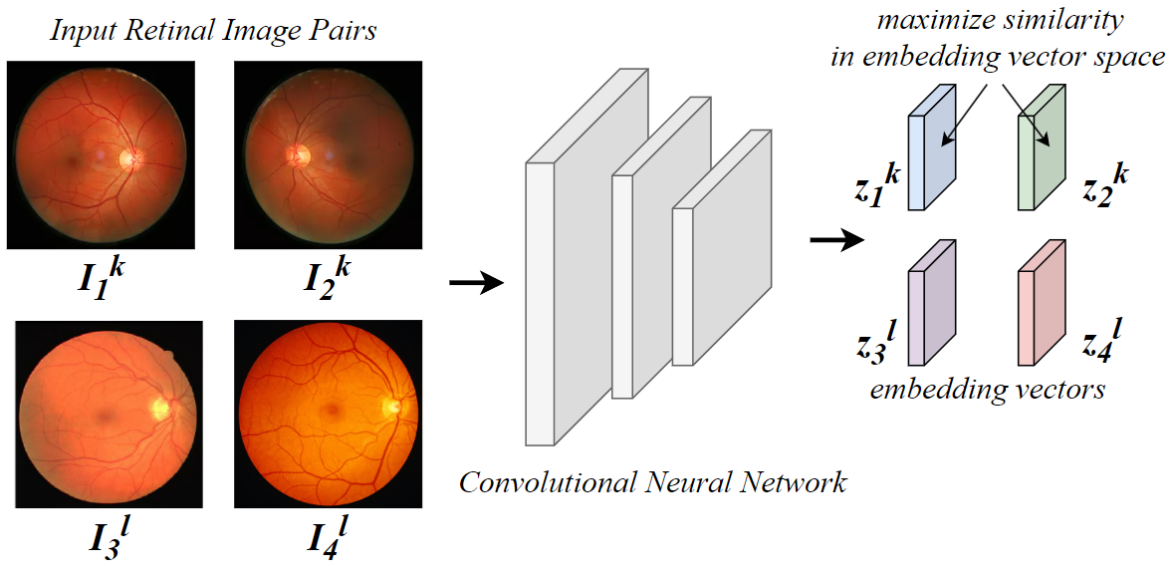

Figure 2. Architecture of the proposed neurodevelopmental disorder diagnosis framework (representation learning part)

Figure 2 depicts the representation learning process proposed in this study. Four retinal images are input to a convolutional neural network, which extracts features or embedding vectors. Each embedding vector captures meaningful characteristics inherent to the input retinal images. The objective of this representation learning phase is to maximize the similarity between two embedding vectors extracted from the same category.

The convolutional neural network takes the four pairs of retinal images I_1^k, I_2^k, I_3^l, and I_4^l as input and produces its corresponding embedding vectors Z_1^k, Z_2^k, Z_3^l, and Z_4^l. Note that, I_1^k and I_2^k are in the same category and I_3^l and I_4^l are in the same category, either Autism spectrum disorder (ASD) or attention-deficit/hyperactivity disorder (ADHD). The fundamental assumption is that when a pair of two retinal images belong to the same category, their embedding vectors should exhibit mathematical similarity. The goal of the proposed representation learning is to maximize this similarity among the embedding vectors extracted from retinal images within the same category.

To measure the similarity between two embedding vectors, we introduce the cosine similarity function as shown in Equation 1.

Equation 1: Cosine Similarity Function

$$S_{i,j} = \frac{z_i^k \circ z_j^k}{|z_i^k| \times |z_j^k|}$$

In the equation 1, $S_{i,j}$ denotes the similarity score between embedding vectors Z_i and Z_j. The operation ∘represents the dot product. If the two input vectors have identical directions, the similarity score becomes 1; otherwise, if the directions are opposite, the result becomes -1. Given this characteristic, we can utilize these calculated similarities as scores for implementing a classification training approach. We now use these similarities to compute probabilities for each similarity score using the softmax function. Equation 2 provides an explanation of how the softmax function is calculated.

Equation 2: Softmax Function

$$P_{i,j} = \frac{e^{S_{i,j}}}{\sum_{m=1}^{M} e^{S_{i,m}}}$$

In equation 2, $P_{i,j}$ denotes the probability of similarity score $S_{i,j}$. The converted probability is then used to measure the error using the cross-entropy loss function (Mao et al. 2023). The cross-entropy loss function is explained in Equation 3.

Equation 3: Cross-entropy Loss Function (loss of similarity)

$$L_{cross-entropy-similarity} = -log_e P_{i,j}$$

As an example, when the similarity score $S_{1,2}$ is maximized, the probability $P_{1,2}$ approaches 1, resulting in a loss value of 0, which represents the ideal scenario. Conversely, if the similarity score $S_{1,2}$ is minimized, $P_{1,2}$ approaches 0, leading to an infinite loss value, which is considered the worst-case scenario. During the representation learning, the convolutional neural network learns to consistently extract similar embedding vectors from retinal images belonging to the same category. This training strategy enhances the network's generalization capabilities for retinal images, resulting in improved classification accuracy during subsequent neurodevelopmental disorder diagnosis tasks.

For the architecture of the convolutional neural network, we chose the ResNet-50 (He et al. 2016) structure, which has demonstrated robust performance across various image classification tasks. We conducted a comprehensive set of experiments, testing multiple convolutional neural network architectures, including VGG19 (Simonyan et al. 2014), MobileNetV2 (Sandler et al. 2018), EfficientNet-B7 (Tan et al. 2019), and HRNet-w32 (Wang et al. 2020).

Fine-Tuning for Downstream Task

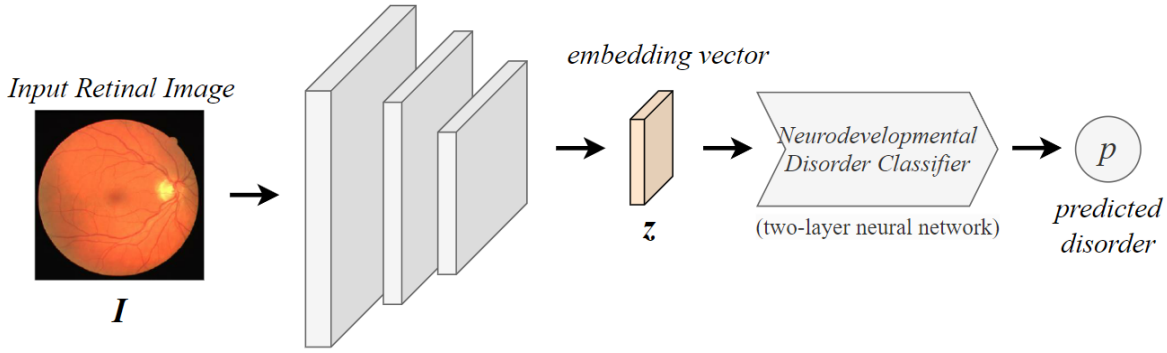

Figure 3. Architecture of the proposed neurodevelopmental disorder diagnosis framework (fine-tuning part)

Figure 3 illustrates the overall architecture of the transfer learning and fine-tuning approach employed in the proposed system. During the fine-tuning process, a single retinal image is input into the convolutional neural network to generate an embedding vector. This extracted embedding vector is subsequently passed to the neurodevelopmental disorder diagnosis classifier, which predicts the type of disorder presented in the inputted retinal image. Notably, the pre-trained convolutional neural network used in the previous representation learning phase serves as the starting point for the fine-tuning process, rather than initializing random variables. This approach accelerates the training process and enhances the accuracy of the diagnosis classification.

To measure the error between the predicted disorder and its ground truth, we simply utilized the cross-entropy loss function which is often used in many classification problems. Equations 4 and 5 explain the mathematical form of the cross-entropy loss function for the disorder and gender classier.

Equation 4: Cross-entropy Loss Function (loss of predicted disorder type)

$$L_{cross-entropy-disorder} = -log_e P$$

Equation 5: Cross-entropy Loss Function (loss of predicted gender)

$$L_{cross-entropy-gender} = -log_e g$$

In the equation 4, 5 P and g denote the probability of the predicted disorder type and gender, respectively.

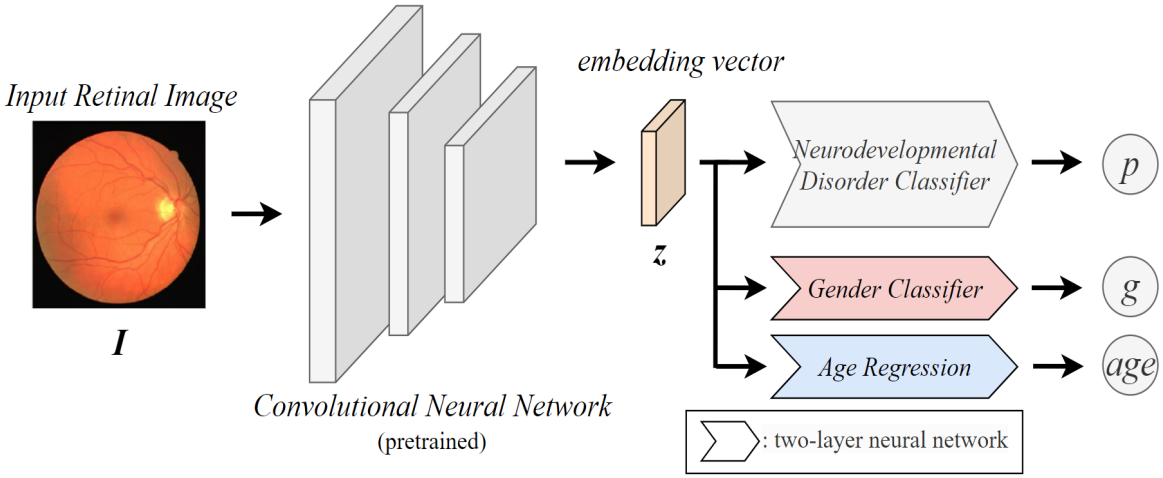

Figure 4. Architecture of the proposed neurodevelopmental disorder diagnosis framework (fine-tuning part, with auxiliary gender classifier and age regression network)

Additionally, we introduce an age and gender-guided training approach, as shown in Figure 4. Age and gender exhibit correlations with neurodevelopmental disorders. To leverage this characteristic within the proposed machine learning system, we conduct joint training for a gender classification network and an age regression network. For assessing the accuracy of predicted gender and age, we employ the cross-entropy loss function and the mean squared error function, respectively. Equation 5 provides an explanation of the mean squared error function.

Equation 5: Mean Squared Error Function

$$L_{mean-squared-error} = \frac{1}{N}\sum_{n=1}^{N}|y - \hat{y}|^2$$

Here $\hat{y}$ and y represent the predicted age and its ground truth age. The variable N denotes the total number of samples. The mean squared error function simply measures the average error between the prediction and its ground truth. Finally, the total loss function is explained in equation 6.

Equation 6: Total Loss Function (fine-tuning)

$$L_{mean-squared-error} = L_{cross-entropy-disorder} + \alpha L_{cross-entropy-gender} + \beta L_{mean-squared-error}$$

Here α and β weights for each gender classification and age regression loss. Through extensive experimentation, we have found that setting the weight for gender classification (α) to 0.6 and the weight for age regression (β) to 0.8 yields the optimal results.

For the architecture design, we utilized two-layer neural networks for both the gender classification and age regression networks, as well as the neurodevelopmental disorder classifier. We found that increasing the depth of the neural network beyond two layers did not significantly contribute to improving accuracy. Therefore, we determined that two layers were sufficient to achieve state-of-the-art performance.

Experimental Results

Dataset

In this chapter, we provide an overview of the dataset used in this research study. The dataset comprises a comprehensive collection of retinal samples obtained from a total of 6,146 patients. It includes a total of 412,559 retinal image samples labeled into three categories: normal, ADHD, and ASD. The normal label contains 329,259 samples, constituting 79.81% of the total samples in the dataset. In the ADHD label, there are 63,250 samples, representing 15.33% of the total samples. Finally, ASD comprises 20,050 samples, making up 4.86% of the total samples.

The age groups are divided as follows. Age under 7: approximately 27.70% of the patients fall into this age category, reflecting a significant portion of the dataset. Age Between 7 and 13: A substantial 43.57% of the patients are within this age range and Age Between 13 and 21: the remaining 28.73% of patients.

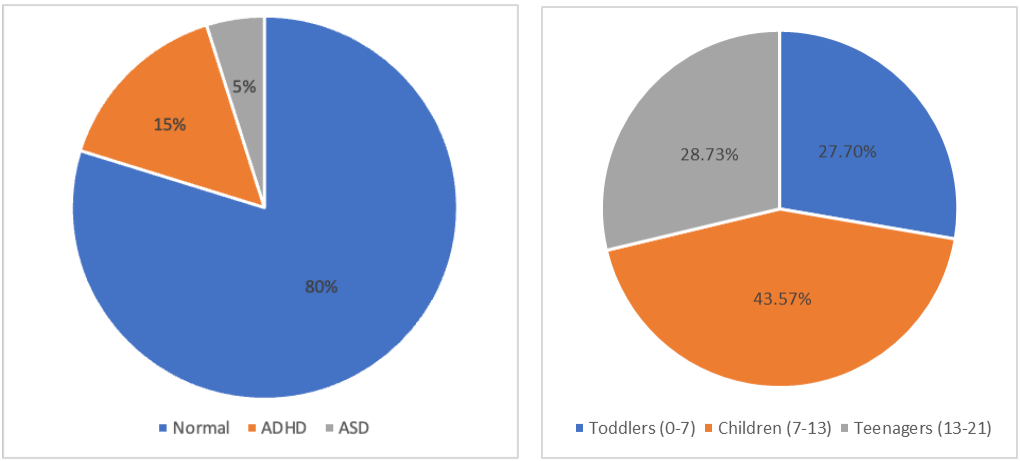

Figure 5. Sample distribution of the dataset used in this research

Experimental Procedure

To assess the effectiveness of the proposed method, we conducted a comprehensive series of experiments, which encompassed performance comparisons, the generation of confusion matrices and receiver operating characteristic curves, the exploration of various network architectures, and the application of data augmentation

techniques. For the training and evaluation processes, we employed a 5-fold cross-validation approach, a commonly used practice in machine learning research. To ensure fairness and consistency in the comparison procedure, we maintained identical training and testing procedures across all experiments.

Comparison with State-Of-The-Art Methods (ADHD)

Firstly, we compared the performance of the proposed method against several benchmark approaches for the task of ADHD classification. Table 1 and Figure 6 summarize the results of these experiments. For comparison, we selected well-established architectures in the computer vision domain, namely VGG19, MobileNetV2, EfficientNet-B7, HRNet-32, and ResNet-50.

Table 1. Performance comparison (ADHD classification task)

	Accuracy	**Recall**	**Precision**	**F1-Score**
VGG19 (Simonyan et al. 2014)	0.8695 (±0.0007)	0.8407 (±0.0008)	0.7122 (±0.0011)	0.7676 (±0.0008)
MobileNetV2 (Sandler et al. 2018)	0.8723 (±0.0005)	0.8578 (±0.0014)	0.7085 (±0.0005)	0.7762 (±0.0011)
EfficientNet-B7 (Tan et al. 2019)	0.8806 (±0.0011)	0.8662 (±0.0007)	0.7194 (±0.0007)	0.7804 (±0.0014)
HRNet-32 (Wang et al. 2020)	0.9023 (±0.0012)	0.8785 (±0.0007)	0.7487 (±0.0006)	0.7985 (±0.0008)
Resnet-50 (He et al. 2016)	0.9084 (±0.0012)	0.8874 (±0.0008)	0.7542 (±0.0011)	0.8095 (±0.0012)
Proposed Method (Resnet-50 based)	**0.9452 (±0.0008)**	**0.9200 (±0.0010)**	**0.7795 (±0.0001)**	**0.8439 (±0.0008)**

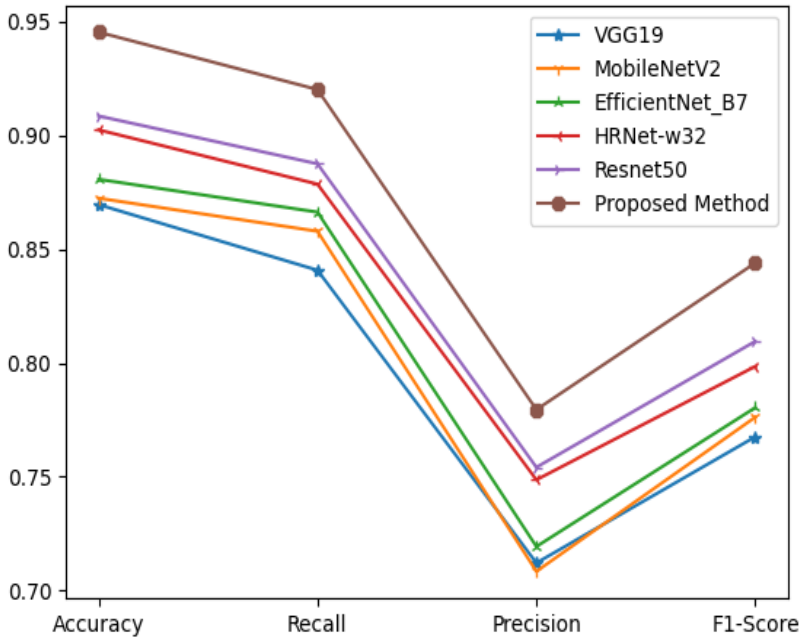

Figure 6. Performance comparison (ADHD classification task)

Among these networks, VGG19, MobileNetV2, and EfficientNet-B7 displayed relatively poor results in the diagnosis of ADHD. This can be attributed to their relatively shallow convolutional layers, which hindered their ability to extract meaningful visual characteristics from the retinal images. Conversely, HRNet-32 and ResNet-50, both equipped with deeper convolutional layers, exhibited slightly improved performance.

Ultimately, the proposed self-supervised-based method outperformed all the supervised-based comparison methods by a significant margin. This can be attributed to the self-supervised-based representation approach, which compelled the trained network to extract more robust and consistent features, ultimately leading to superior performance.

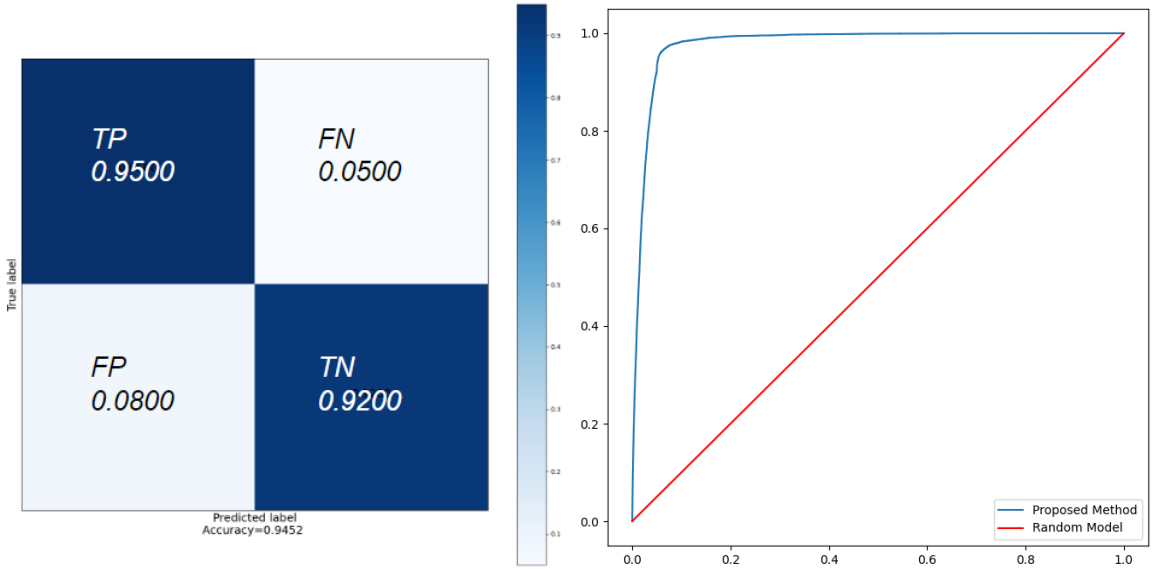

Figure 7. Confusion matrix and ROC (Receiver Operating Characteristic) curve of the proposed method (ADHD classification task) (left): confusion matrix, and (right): ROC curve

Figure 7 presents both the confusion matrix and the receiver operating characteristic curve for the proposed methods in the ADHD classification task. Examining the diagonal components of the confusion matrix provides a clear demonstration of the robustness and consistency achieved by the proposed method. Additionally, the receiver operating characteristic curve assumes an almost rectangular shape in the upper-left corner, indicating that the proposed method exhibits exceptional discriminatory power and the ability to effectively distinguish between positive and negative cases.

Comparison with State-Of-The-Art Methods (ASD)

In the second experiment, we conducted the comparison on the ASD classification task. The experimental protocol remains identical to the one detailed in Chapter 4.3.

Table 2. Performance comparison (ASD classification task)

	Accuracy	Recall	Precision	F1-Score
VGG19	0.8983 (±0.0011)	0.8333 (±0.0013)	0.6629 (±0.0009)	0.7308 (±0.0010)
MobileNetV2	0.9062 (±0.0009)	0.8311 (±0.0011)	0.6655 (±0.0010)	0.7407 (±0.0009)
EfficientNet-b7	0.9183 (±0.0014)	0.8427 (±0.0011)	0.6727 (±0.0013)	0.7486 (±0.0008)
HRNet-32	0.9208 (±0.0016)	0.8554 (±0.0008)	0.6896 (±0.0009)	0.7545 (±0.0010)
Resnet-50	0.9367 (±0.0007)	0.8522 (±0.0011)	0.6838 (±0.0009)	0.7584 (±0.0016)
Proposed Method (Resnet-50 based)	**0.9754 (±0.0010)**	**0.9004 (±0.0011)**	**0.7325 (±0.0009)**	**0.8077 (±0.0012)**

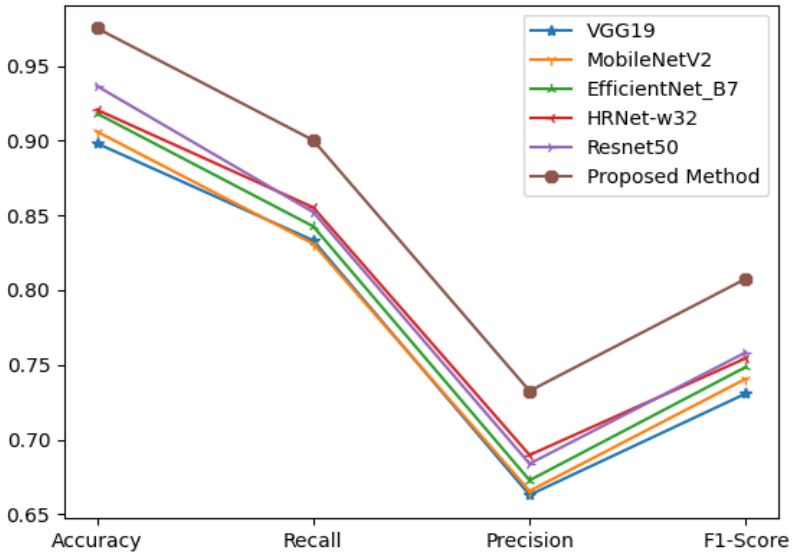

Figure 8. Performance comparison (ASD classification task)

Table 2 and Figure 8 provide a summary of the experimental results. Once again, the shallow convolutional neural networks, specifically VGG19, MobileNet, and EfficientNet-B7, yielded poor results, while the deep networks showed slightly improved performance. Notably, the proposed method consistently outperformed all the compared methods, offering clear evidence of the effectiveness of the self-supervised representation learning approach we have introduced.

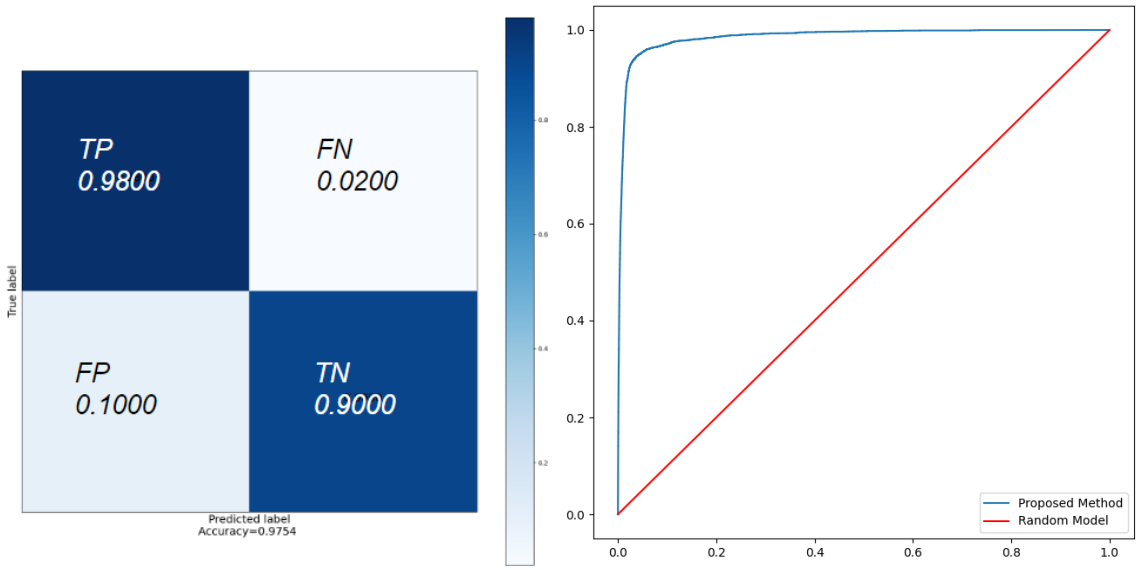

Figure 9. Confusion matrix and ROC curve of the proposed method (ASD classification task) (left): confusion matrix, and (right): ROC curve

Figure 9 displays both the confusion matrix and receiver operating characteristic curve of the proposed method in the ASD classification task. The observed results exhibit a similar trend compared to the ADHD classification task. In the following chapters, we explain additional experiments in order to comprehensively examine the effectiveness of the proposed method from various perspectives.

Ablation Study

In this chapter, we perform an experiment focused on replacing the network architecture. Given that the proposed method follows a plug-and-play approach, we explore the behavior of the method when different convolutional neural network architectures are employed. While maintaining consistent experimental hyperparameters, we exclusively vary the convolutional neural network architectures. Table 3 and Figure 10 provide summaries of the experimental outcomes.

Table 3. Architecture replacement experiment (ASD classification)

	Accuracy (baseline)	**Accuracy (proposed method)**
VGG19	0.8983 (±0.0011)	0.9197 (±0.0007)
MobileNetV2	0.9062 (±0.0009)	0.9295 (±0.0013)
EfficientNet-b7	0.9183 (±0.0014)	0.9461 (±0.0008)
HRNet-32	0.9208 (±0.0016)	0.9552 (±0.0009)
Resnet-50	0.9367 (±0.0007)	0.9754 (±0.0010)

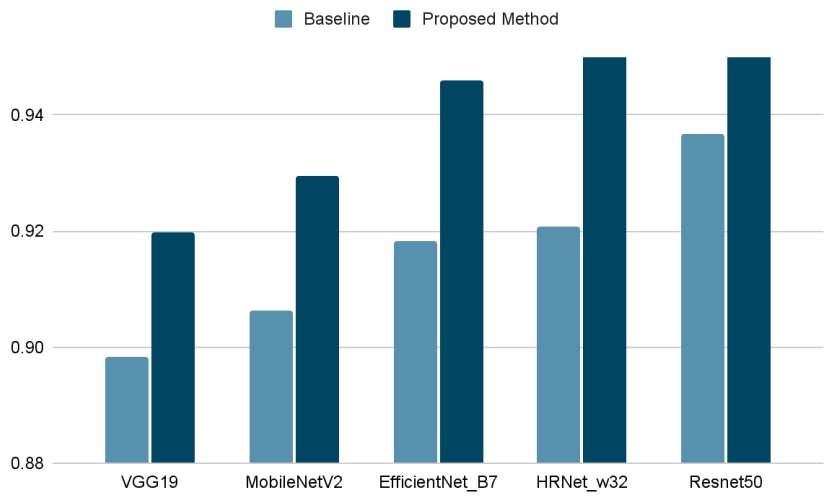

Figure 10. Architecture replacement experiment (ASD classification)

Remarkably, the application of the proposed method to various network architectures yielded increased accuracy, spanning both shallow and deep convolutional neural networks. This outcome clearly illustrates that the proposed approach is not biased toward a specific architecture but can be effectively applied and generalized across a wide range of network architectures.

Additional Experiments

Finally, we explore several data augmentation techniques to enhance the model's accuracy further. In this experiment, we initially trained the network without any augmentation, which serves as our baseline. Subsequently, we applied four augmentation methods, including grayscale conversion, YCbCr color space transformation, sharpening, and histogram equalization. Each data augmentation method is then compared to the baseline to illustrate its impact on model performance.

Table 4. Data augmentation experiment

	Accuracy (ADHD)	**Accuracy (ASD)**
Baseline	0.9452 (±0.0008)	0.9754 (±0.0010)
Grayscale	0.9075 (±0.0007)	0.9471 (±0.0008)
YCbCr	0.9253 (±0.0011)	0.9623 (±0.0009)
Sharpness	0.9187 (±0.0013)	0.9575 (±0.0011)
Histogram Equalization	0.9589 (±0.0009)	0.9862 (±0.0007)

Table 4 provides a summary of the experimental results. Grayscale, YCbCr, and sharpness augmentation did not lead to an increase in accuracy, whereas histogram equalization resulted in improved accuracy. This can be attributed to the fact that histogram equalization ensures a consistent pixel range distribution in the input retinal images, potentially enhancing the stability and accuracy of the inference process.

Conclusion

In this research study, we proposed a self-supervised representation learning for neurodevelopmental disorder diagnosis system. The proposed system takes retinal images as input and outputs the predicted diagnosis category. The proposed method studied the potential of deep learning and self-supervised representation learning in enhancing the accuracy and robustness of neurodevelopmental disorder diagnosis. Through extensive experiments, we have proved that the proposed self-supervised representation learning approach consistently outperformed supervised-based methods, demonstrating its effectiveness in capturing more robust and consistent features from retinal images. The proposed method exhibited remarkable versatility, performing well across various convolutional neural network architectures. Additionally, the application of data augmentation techniques, particularly histogram equalization, led to improved accuracy. In the future, we plan to expand the proposed method to other neuro-related diseases and disorders.

Acknowledgments

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

References

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

Mao, A., Mohri, M., & Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications. arXiv preprint arXiv:2304.07288. https://doi.org/10.48550/arXiv.2304.07288

Patton, N., Aslam, T. M., MacGillivray, T., Deary, I. J., Dhillon, B., Eikelboom, R. H., ... & Constable, I. J. (2006). Retinal image analysis: concepts, applications and potential. Progress in retinal and eye research, 25(1), 99-127.

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

Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR. https://doi.org/10.48550/arXiv.1905.11946

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

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., ... & Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. Nature, 1-8.