A 3D-ResNet Combined with BRNN: Application in the Auxiliary Diagnosis of ADHD

Abstract

Attention Deficit/Hyperactivity Disorder (ADHD) is a common mental disorder that exhibits a high incidence rate in children and adolescents, and it is also observed in adults. Currently, there is a lack of objective diagnostic methods for ADHD. Therefore, a three-dimensional residual network (3D-ResNet) deep learning method based on feature extraction from rs-fMRI images for assisting in the diagnosis of ADHD based on resting-state functional magnetic resonance imaging (rs-fMRI) and deep learning models was proposed in this paper. Taking into consideration the temporal characteristics of rs-fMRI, we constructed a 3D-ResNet model based on four-dimensional image. The model utilized TimeDistributed to encapsulate residual blocks which allowed the model to extract spatial features from rs-fMRI while preserving its temporal sequence information. We constructed four different hierarchical structures of 3D-ResNet which are subsequently combined with two different bidirectional recurrent neural networks (BRNNs) to extract sequence features. And BRNNs includes bidirectional long short-term memory (Bi-LSTM) and bidirectional gated recurrent unit (Bi-GRU). The proposed method utilized the ADHD-200 Consortium's public dataset for training and was validated by 5-fold cross-validation. The experimental results indicated that the proposed method in this study demonstrated superior performance on the dataset compared to traditional methods (Accuracy: 76.56%, Sensitivity: 80.16%, Specificity: 90.22%). Therefore, adopting this method can further enhance the accuracy of assisting in the diagnosis of ADHD.

1 Introduction

Attention Deficit/Hyperactivity Disorder (ADHD) is a highly prevalent neurodevelopmental disorder[1,] the onset of the condition typically occurs before the age of 12[2], and it is characterized by persistent hyperactivity, excessive impulsivity, or inability to concentrate[3]. ADHD can be classified into three subtypes, namely Attention Deficit/Hyperactivity Disorder-Combined Type (ADHD-C), Hyperactive-Impulsive Type (ADHD-HI), and Inattentive Type (ADHD-I).

To mitigate the challenges associated with image fusion, this study constructed a deep learning model based on rs-fMRI image data. The paper introduced a network architecture named 3D-ResNet, which is employed to extract spatial features from rs-fMRI. Subsequently, it combines Bidirectional Recurrent Neural Networks (BRNNs) to extract temporal features. Traditional Recurrent Neural Networks (RNNs) face challenges such as gradient vanishing and exploding when dealing with sequential data17, these problems make it difficult to capture long-range dependencies. BRNNs capture dependency relationships in sequential data by combining information from both the forward and backward directions. Unlike RNNs, which only consider past information, BRNNs simultaneously take into account both past and future information[3]. This ability helps the model to comprehensively understand the context within the sequence and consequently improve its classification accuracy. The paper combines two different types of BRNN—Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU)—to find the optimal network composition, Bi-LSTM is an extended form of Long Short-Term Memory (LSTM)[4], and Bi-GRU is a bidirectional recurrent neural network based on Gated Recurrent Unit (GRU)[5]. Unlike LSTM and GRU, Bi-LSTM and Bi-GRU can simultaneously consider past and future information at each time step, and that enables better capture of long-term dependencies in time series. Compared to traditional methods, the main contributions of our method are as follows:

Using scaled rs-fMRI image data as input avoids the cumbersome preprocessing steps associated with multimodal image fusion, and it reduces the need for extensive feature engineering.

A deep learning model based on four-dimensional image data was innovatively constructed. This model focuses on extracting spatial features while preserving fMRI time series, and that enhances the correlation of features in both spatial and temporal dimensions.

Four different structures of 3D-ResNet were designed, and they were combined with two types of BRNNs. Through ablation experiments, the optimal model combination was identified.

2 Methods

The model proposed in this paper takes rs-fMRI image with a temporal dimension as input. It extracts spatial and temporal features through different modules. The model consists of three modules: spatial feature extraction network, temporal feature extraction network, and classifier. Due to rs-fMRI being a four-dimensional image with a temporal dimension, to preserve its inherent temporal characteristics when extracting the spatial features of rs-fMRI images, this paper innovatively builds a three-dimensional residual network (3D-ResNet) based on three-dimensional convolutional neural networks (3D-CNNs). This is a sequential combination network that simultaneously considers spatial and temporal features. The 3D-ResNet network is utilized to extract spatial features while preserving the temporal features of the images. Subsequently, the extracted feature sequence is used as input for the second module, where a temporal feature extraction network processes the sequence. Finally, the obtained features are fed into a classifier. The following three sections will provide detailed compositions of each module.

3 Experiment

3.1 Dataset and preprocessing

This paper trains on rs-fMRI data that has been preprocessed using the Athena pipeline. The samples are sourced from the ADHD-200 Global Competition dataset. The Athena pipeline provides information on 973 preprocessed subjects, including rs-fMRI scans, T1-weighted structural scans, and preprocessed script files. The preprocessing steps primarily include operations such as slice timing correction, head motion correction, smoothing, and filtering. To mitigate the impact of age differences and the imbalanced distribution of positive and negative samples on model training, after stage exclusion, the remaining rs-fMRI data from 430 subjects is used as input. ADHD subtypes are ignored, and all subclasses are labeled as 1. The average age of participants is 12.62, with an equal proportion of ADHD to Typically Developing Control (TDC) subjects at a ratio of 1:1. For a detailed composition of the dataset, refer to Table 1.

Table 1: The detailed composition of multi-site samples.

	Pittsburgh	Peking	Total
ADHD	0	78	215
TDC	49	88	215
Total	49	166	430

After preprocessing, the spatial dimensions of the rs-fMRI data are 49x58x47. However, due to variations across different sites, the length of the time series is not uniform. For example, the fMRI data from the NYU site has a scan time of 172, while the image time series length from the NeuroIMAGE site is 257. To mitigate the potential impact of differences in scanners and parameters across different sites on experimental results, the original images are cropped and resized to 20x34x34x34x1 before training. This size is then used as the input dimension for the model, where 20 represents the length of the rs-fMRI time series, and 34 represents the spatial dimensions in terms of length (h), width (w), and depth (d). This approach not only standardizes the image sizes across different sites but also substantially reduces the model's parameter count, simultaneously it can avoid overfitting risks and reduce memory overhead[6-8].

3.2 Model training

In this study, all models are trained using binary cross-entropy loss function with the Adam optimizer. The learning rate is set to 1x10-4. Given that fMRI images are four-dimensional with relatively large dimensions, the images are scaled proportionally before training. The standardized size after scaling is 20x34x34x34x1. Subsequently, the samples are fed into the model for training. The batch size is set to 32, and the number of epochs is set to 100. The model's training progress is evaluated using the accuracy metric. After multiple parameter adjustments, the training proceeds successfully.

To avoid overfitting during the training process, this study employs the early stop technique with a tolerance set to 5. In other words, if the loss does not show a decreasing trend for five consecutive iterations, the training is terminated, and the model parameters from five iterations ago are saved. Since this paper proposes multiple networks with different structures, to comprehensively evaluate the model performance, multiple metrics are incorporated in the model evaluation. Additionally, a five-fold cross-validation is used to obtain more accurate classification performance.

4 Results

This section primarily presents the experimental results of training four different residual networks combined with two types of BRNNs, and compares them with existing models. In certain research reports focusing on classification tasks, the majority often rely solely on the accuracy metric to assess their methods. However, this alone is inadequate to substantiate the feasibility of their approaches, that's because high classification accuracy may be a result of imbalanced sample distribution, and it will lead the model to exhibit bias towards predicting a specific class in extreme cases. For instance, in a binary classification task where positive samples constitute only 10% of the entire dataset, if the model predicts all samples as negative, the accuracy can reach 90%. However, for the 10% positive samples, the model's ability to accurately predict is uncertain. In this case, the high accuracy is superficial and lacks practical significance. Therefore, to accurately assess the model performance, this paper introduces specificity and sensitivity. Specificity represents the false positive rate, and a high specificity indicates a low number of misdiagnosed samples. In simple terms, it reflects the model's ability to correctly identify TDC. It can be calculated as follows:

$$specificity = \frac{TN}{TN + FP}$$

Sensitivity means the proportion of samples not missed in the model's prediction results relative to the total number of samples. It represents the model's ability to correctly identify ADHD cases. The calculation for sensitivity is as follows:

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

Where True Negative (TN) represents the number of samples that are negative and predicted as negative, True Positive (TP) represents the number of samples that are positive and predicted as positive, False Negative (FN) represents the number of samples that are positive but predicted as negative, and False Positive (FP) represents the number of samples that are negative but predicted as positive[9].

4.1 Results of combining Residual Networks with Bi-LSTM

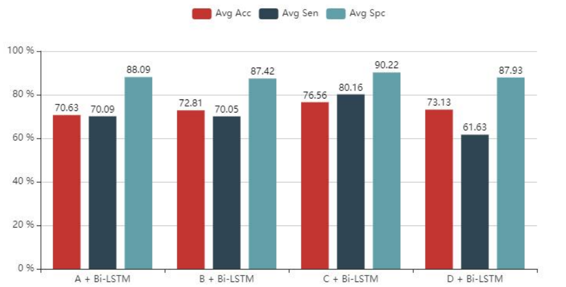

Figure 1: The experimental results of the combination of 3D-ResNet and Bidirectional LSTM model

In Figure 1, the experimental results of training with the combination of four Residual Networks and Bi-LSTM are presented. The data in the table represents the average results as the paper utilized five-fold cross-entropy validation. From the experimental results, it can be observed that the performance of these four combined models is quite close, with the main differences manifesting in terms of accuracy and sensitivity.

In the case of combining with Bi-LSTM, the accuracy of Residual Networks A and B is inferior to that of Residual Networks C and D. The combined model of Residual Network C achieved the highest sensitivity of 80.16% and the highest accuracy of 76.56%. During the training process, the models of Residual Networks C and D exhibit a faster convergence rate compared to those of Residual Networks A and B. Figure 5 illustrates the performance of this model using 5-fold cross-entropy validation. As shown in the figure, except for validation set 4, the accuracy of other validation sets is greater than 75%. The comprehensive performance of sensitivity and specificity indicates that this combination has good adaptability and can fit the model's classification curve well.

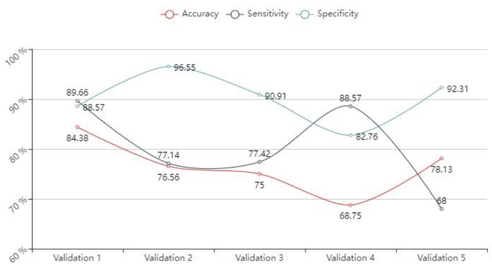

Figure 2: The performance of the combination of Residual Network C and Bi-LSTM using 5-fold cross-entropy validation

4.2 Results of combining Residual Networks with Bi-GRU

Following the same method as the previous section, four Residual Networks were individually combined with Bi-GRU for training and validation. Figure 3 illustrates the performance of the models after 5-fold cross-entropy validation. The main difference in the current combination method is reflected in sensitivity, while the four Residual Networks show similar performance in accuracy and specificity. From the figure, it can be observed that the combined model of Residual Network C has the best overall performance across these three metrics, and its accuracy, sensitivity, and specificity are 71.25%, 70.97%, and 89.13%. When the model's accuracy and specificity exhibit similar performance, sensitivity becomes a key indicator representing the performance differences among the four models. The performance of the model on the validation set at this time is shown in Figure 4.

The data in the figure indicates that except for validation set 2, the accuracy of the model combining Residual Network C with Bi-GRU is less than 75%. The accuracy on validation sets 4 and 5 is even less than 70%. Compared to the model combined with Bi-LSTM, the combination of Residual Network C and Bi-GRU is not stable enough, especially in terms of sensitivity. It can't fit the classification curve well and only performs well in certain specific intervals, and the results lack generality. Therefore, by comparing the data, it can be concluded that the model combining Residual Network C with Bi-LSTM exhibits the best performance in the task of ADHD classification recognition.

Figure 3: The experimental results of the combination of 3D-ResNet and Bidirectional GRU model

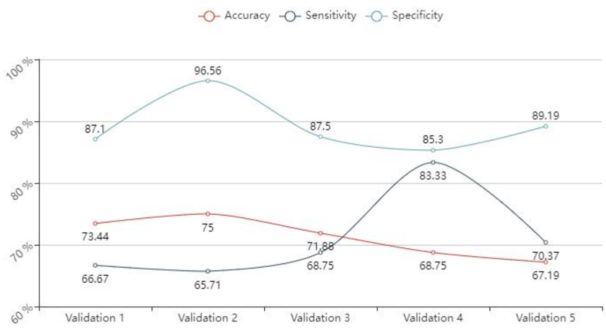

Figure 4: The performance of the combination of Residual Network C and Bi-GRU using 5-fold cross-entropy validation

4.3 Compare with existing models

Table 2 compares various existing methods with the approach proposed in this paper. Firstly, all of them are based on the ADHD-200 dataset, it can be observed that the proposed method in this paper shows a significant improvement compared to existing methods, both in terms of accuracy, sensitivity, and specificity. However, due to variations in data partitioning across different studies, and the fact that some studies do not report specificity and sensitivity, so making direct comparisons is challenging.

Table 2:Comparison of performance metrics across models.

	Methods	Validation	Accuracy	Sensitivity	Specificity
Dai et al.[7]	MKL	10-fold cross	67.79%	38.29%	84.08%
Zou et al.[11]	3D-CNN	hold-out set[1]	69.15%	-	-
Mao et al.[13]	4D-CNN	hold-out set[1]	71.30%	73.20%	69.70%
Zhang et al.[14]	SC-CNN-Attention	loocv[2]	68.60%	-	-
Niu et al.[16]	3D-CNN+GRU	5-fold cross	71.65%	68.00%	73.80%
Proposed	3D-ResNet+Bi-LSTM	5-fold cross	76.56%	80.16%	90.22%

5 Discussion

In existing research, the majority of studies adopt dimensionality reduction to extract low-order features from fMRI images. This machine learning approach, by employing straightforward feature selection to discard irrelevant features13, and it often overlooks the temporal and spatial correlations in fMRI data. Consequently, the features extracted lack representational capacity. Recent studies tend to lean towards employing deep learning methods for feature extraction, for example, Niu et al.[10] utilized 3D-CNN to process the three-dimensional spatial information of rs-fMRI. Through one-dimensional filters of different scales, significant features were extracted. Subsequently, an independent GRU was employed to handle the one-dimensional temporal information. Finally, feature fusion was performed. Hong et al.[11] employed TimeDistributed to encapsulate 3D-CNN for processing the spatiotemporal information of rs-fMRI. Through this network, a feature sequence was obtained. Then, it was combined with LSTM to extract sequence features, this method achieved a classification accuracy of 68% for ADHD. The method of using 3D-CNN and GRU to separately process the spatiotemporal information of rs-fMRI takes into account the correlation between spatial and temporal dimensions. However, in reality, fMRI images exhibit dynamic spatial characteristics with temporal features. The approach of using two independent networks to extract features separately may contradict the realistic nature of the data[12]. This results in lower spatiotemporal correlation of features. The concatenated network of 3D-CNN combined with LSTM which is the inspiration behind this paper avoided this issue. Due to the use of TimeDistributed encapsulation for 3D-CNN with four-dimensional input data, the model generated a large number of parameters. Setting a high number of convolutional layers may lead to overfitting during the training process. On the other hand, a shallower 3D-CNN may not effectively extract meaningful information, and it will lead to a lack of expressive power in the model and consequently lower accuracy.

In this study, a 3D-ResNet encapsulated by TimeDistributed was constructed to process the spatial information of rs-fMRI. It is combined with BRNNs to form a concatenated network, and that effectively addressed the issues of insufficient spatiotemporal correlation of features and overfitting. By analyzing the experimental results in the fourth section, it can be concluded that the overall performance of the model combining 3D-ResNet with Bi-LSTM is better than that with Bi-GRU. Among them, the model combining Residual Network C with Bi-LSTM exhibits the best performance. However, it is essential to note that this model requires a certain amount of memory as a basis. The experimental model was trained on an NVIDIA A100 GPU with 40GB of memory. Despite the increased memory overhead, the noticeable improvement in accuracy justifies the additional memory usage.

6 Conclusion

A three-dimensional residual network named 3D-ResNet which was combined with BRNNs was introduced in this study. Compared to techniques that involve fusing fMRI with MRI, the method proposed in this paper eliminated the need for complex image preprocessing; And compared to methods that extract low-level features from fMRI, this model retained spatial correlations while extracting features. This paper constructed four different structures of residual networks, and through ablation experiment, it demonstrated that the model combining Residual Network C with Bi-LSTM has the best performance. Under the 5-

fold cross-entropy validation method, the average accuracy, sensitivity, and specificity are 76.56%, 80.16%, and 90.22%. Compared to existing methods, there is a significant improvement in accuracy when performing classification tasks on the multi-site ADHD-200 dataset. This result indicated that combining 3D-ResNet with BRNNs for assisting in the diagnosis of ADHD is feasible. What is even more promising is that this technology can be applied to the classification and diagnosis of other neurological disorders. It holds considerable prospects in studies based on rs-fMRI.

References

Zhou, Dingfu et al. “Deep Learning Enabled Diagnosis of Children's ADHD Based on the Big Data of Video Screen Long-Range EEG.” Journal of healthcare engineering vol. 2022 5222136. 4 Apr. 2022.

Cherkasova, Mariya V et al. “Review: Adult Outcome as Seen Through Controlled Prospective Follow-up Studies of Children With Attention-Deficit/Hyperactivity Disorder Followed Into Adulthood.” Journal of the American Academy of Child and Adolescent Psychiatry vol. 61,3 (2022): 378-391.

Bellec, Pierre et al. “The Neuro Bureau ADHD-200 Preprocessed repository.” NeuroImage vol. 144,Pt B (2017): 275-286.

ADHD-200-Results-Webpage. "Adhd-200 global competition results" 2011. Available online at: http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html.

Dai, Dai et al. “Classification of ADHD children through multimodal magnetic resonance imaging.” Frontiers in Systems Neuroscience 6 (2012): n. pag.

Guo, Xiaojiao and Lianghua He. “ADHD Discrimination Based on Social Network.” 2014 International Conference on Cloud Computing and Big Data (2014): 55-61.

Kuang, Deping et al. “Discrimination of ADHD Based on fMRI Data with Deep Belief Network.” International Conference on Intelligent Computing (2014).

Hao, A.J.; He, B.L.; Yin, C.H.: 'Discrimination of ADHD children based on Deep Bayesian Network', IET Conference Proceedings, 2015, p. 6 .-6 .

Zou, Liang et al. “3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI.” IEEE Access 5 (2017): 23626-23636.

Vu, Hanh et al. “3D convolutional neural network for feature extraction and classification of fMRI volumes.” 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI) (2018): 1-4.

Mao, Zhenyu et al. “Spatio-temporal deep learning method for ADHD fMRI classification.” Inf. Sci. 499 (2019): 1-11.

Zhang, Tao et al. “Separated Channel Attention Convolutional Neural Network (SC-CNN-Attention) to Identify ADHD in Multi-Site Rs-fMRI Dataset.” Entropy (Basel, Switzerland) vol. 22,8 893. 14 Aug. 2020.

Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.