

Auxiliary diagnostic system for ADHD in children based on AI technology*

Yanyi ZHANG^{†1}, Ming KONG^{†2}, Tianqi ZHAO², Wenchen HONG²,
Di XIE³, Chunmao WANG³, Rongwang YANG¹, Rong LI¹, Qiang ZHU^{†‡2}

¹Department of Psychology, The Children's Hospital, Zhejiang University School of Medicine,
National Clinical Research Center for Child Health, Hangzhou 310052, China

²College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

³Hikvision Research Institute, Hangzhou 310052, China

[†]E-mail: doczyy1981@sina.com; zjukongming@zju.edu.cn; zhuq@zju.edu.cn

Received Dec. 25, 2019; Revision accepted June 27, 2020; Crosschecked Nov. 13, 2020

Abstract: Traditional diagnosis of attention deficit hyperactivity disorder (ADHD) in children is primarily through a questionnaire filled out by parents/teachers and clinical observations by doctors. It is inefficient and heavily depends on the doctor's level of experience. In this paper, we integrate artificial intelligence (AI) technology into a software-hardware coordinated system to make ADHD diagnosis more efficient. Together with the intelligent analysis module, the camera group will collect the eye focus, facial expression, 3D body posture, and other children's information during the completion of the functional test. Then, a multi-modal deep learning model is proposed to classify abnormal behavior fragments of children from the captured videos. In combination with other system modules, standardized diagnostic reports can be automatically generated, including test results, abnormal behavior analysis, diagnostic aid conclusions, and treatment recommendations. This system has participated in clinical diagnosis in Department of Psychology, The Children's Hospital, Zhejiang University School of Medicine, and has been accepted and praised by doctors and patients.

Key words: Attention deficit hyperactivity disorder (ADHD); Auxiliary diagnosis; Computer vision; Deep learning; BERT

<https://doi.org/10.1631/FITEE.1900729>

CLC number: TP391.4


1 Introduction

The most severe neurodevelopmental disorder in children and adolescents is attention deficit hyperactivity disorder (ADHD). The clinical symptoms include difficulty in concentration, excessive activity, emotional instability, and learning difficulties. The disease leads to multiple functional impairments in patients during childhood, such as learning abilities,

social skills, family relationships, and self-esteem. Additionally, the disease continues to affect adolescents and adults (Willcutt et al., 2012; Polanczyk et al., 2014; Sayal et al., 2018). Moreover, other studies showed a global ADHD incidence of about 5%, where the prevalence of ADHD in China is about 6.26%, which is higher than the international level, with about 23 million patients in total (Wang et al., 2017). A retrospective study conducted by the World Health Organization (WHO) in 10 countries found that ADHD has a 50% adult prevalence rate and is more likely to suffer from various adult mental illnesses. For example, patients with ADHD are more likely to have comorbid substance abuse,

[‡] Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61625107)

 ORCID: Yanyi ZHANG, <https://orcid.org/0000-0001-5238-1712>; Ming KONG, <https://orcid.org/0000-0002-6177-3707>; Qiang ZHU, <https://orcid.org/0000-0002-2405-6776>

© Zhejiang University Press 2021

with a risk of lifetime use of nicotine and illegal substances. About 16%–31% of the clinically confirmed adult ADHD patients also meet the major depressive episode, which is about five times the general population. Among children with ADHD, the rate of early antisocial personality disorder in adults is 18%–24%. It is, therefore, of great medical and social significance to improve the diagnosis ability of ADHD and detect the disease in an early stage.

The standard evaluation and monitoring approach for children's ADHD is to collect proof of a child's everyday actions by parents/teachers in families/schools. The preliminary diagnosis is based on the results of psychology questionnaires, along with the doctors' findings during the consultation period. The conventional approach has the following problems: (1) Because of a lack of medical services, doctors must restrict each child's outpatient time, which sometimes makes doctors unable to interact thoroughly with them. The child is psychologically suppressed by the impact of a fresh atmosphere and authority, so it is difficult to evaluate the child's true actions correctly. (2) Parental concern regarding abnormal behaviors of children and personality differences can seriously affect the objective accuracy of the parental feedback. Due to time and space constraints, the input from the teachers cannot be obtained in real time, so it is either inaccessible or has significant deviations. (3) Because of the diversity and non-standardization of information sources, the information is prone to inconsistencies or even contradictions, making the diagnostic conclusions entirely dependent on the physician's subjective judgment and level of experience. In reality, multiple visits are necessary for diagnosing ADHD accurately, resulting in unnecessary wastage of medical resources and increasing the burden on families.

In this study, we propose an artificial intelligence (AI) based software-hardware cooperation program to assist in the assessment and diagnosis of ADHD in children. This approach incorporates software-based executive functional testing (i.e., completing a task to assess children's inhibition function or cognitive transferability). We then use self-designed hardware modules (consisting of three cameras) to record visual information such as eye motions, facial expressions, and three-dimensional (3D) body postures. This, in effect, feeds into a multi-modal deep learning model (BERT) for detect-

ing abnormal behaviors in children from the recorded video. Finally, we combine the results of various scale questionnaires, the results of executive function evaluations, and the smart detection of children's activities to automatically generate objective, measurable, and regular auxiliary diagnostic reports for doctors and parents' references. Compared to traditional solutions, the three significant innovations and contributions of this research are as follows: (1) the introduction of computer vision technology into the ADHD diagnostic process to assist doctors in the measurable analysis of behavioral characteristics; (2) the use of BERT-based multi-modal fusion technology to assess the time segment of the abnormal behavioral character of the patient; (3) the presentation of a more procedural and standardized ADHD diagnosis mode compared with the traditional diagnostic process. Without raising doctors' workload, it can track the actions of patients more adequately and form a more detailed diagnostic report with diagnostic bases.

2 Related work

AI has been widely used within the medical field in recent years. For example, deep learning uses models of a convolutional neural network (CNN) to complete tasks such as detection, segmentation, and medical image classification. Extensive research of the related techniques has been conducted on various clinical fields, such as skin disorders, diabetic fundus, brain magnetic resonance imaging (MRI), chest X-rays, CT, and pathological cancer cells. Natural language processing (NLP) can analyze unstructured medical record information, extract and structure critical information, which is useful for standardizing and analyzing medical record data.

Compared with issues such as diabetic fundus and X-ray-based lung disease examination, studies on computer-aided diagnosis for mental illnesses, such as ADHD, are few. Zou et al. (2017) continued the benefits of in-depth medical imaging research, and attempted to examine the differences between ADHD patients and healthy children from the brain MRI image perspective. The electroencephalogram (EEG) was used (Marcano et al., 2018) based on the finding that the EEG signals of ADHD patients differ from those of healthy people. Chen et al. (2019) developed a multichannel deep learning network to

analyze patient brain rs-fMRI data for ADHD diagnosis. Aradhya et al. (2019) further captured the spatio-temporal correlation between different brain regions, significantly improved the diagnosis accuracy, and ensured the consistency of model logic and medical logic. Such works much inspired the study of AI-based ADHD analysis. However, these researches' rationalities were not fully confirmed or validated from the medical point of view. Moreover, clinically, these findings still significantly differ from the current method of diagnosing ADHD, which is not conducive to enhancing the actual clinical diagnosis process.

Another type of research is dedicated to examining the visual characteristics such as attention, expression, and movement of ADHD patients (Jaiswal et al., 2017). As our closest research on diagnosing the ADHD and autism spectrum disorder (ASD) with computer vision analysis technology, they used RGBD (color + deep) sensors to collect the facial visual signals of the testee, taking various features into account, such as the facial expression, head position, body movement, distance of motion, and response time, and classified the ADHD and ASD with the support vector machine (SVM). Leo et al. (2018) used the computer vision technology to quantitatively assess children with ASD's ability to produce facial expressions. Muñoz-Organero et al. (2019) used wrist and ankle acceleration sensors to track the child's activity status, and measured the actions of typical children and ADHD patients using the recurrent neural network (RNN) based model. Li et al. (2019) used hierarchical long-short term memory (LSTM) to examine time-series eye movement data from children with ASD to help diagnose the disorder. These researches show that visual perception technology and machine learning technology for intelligent diagnosis of ADHD have feasibility and research value. However, their work was not well integrated with the traditional methods and basis for diagnosing ADHD. Studying pure visual behaviors is not sufficient to support the complete logic chain of ADHD diagnosis, significantly limiting the medical landing value of this solution. Additionally, they did not use the new deep learning models to fuse and discern multidimensional signals, which also hinders their work's technical development.

Our study defines and analyzes children with ADHD from multidimensional perspectives. The

basis for our inspiration comes from the cooperating doctor team's existing diagnostic experience and logic. AI technology captures and analyzes facial expressions, eye focus, movements, and other content, and quantifies those findings using a deep learning fusion model as an objective diagnostic basis. The intelligent auxiliary diagnostic system finally generates a diagnostic report that can be fully understood by the doctors and parents, making the diagnostic solution both feasible and useful.

3 System design

The ADHD intelligent auxiliary diagnostic system is composed mainly of four modules: scale test module, software-hardware coordination module, intelligent analysis module, and multi-modal fusion module. The specific structure is shown in Fig. 1.

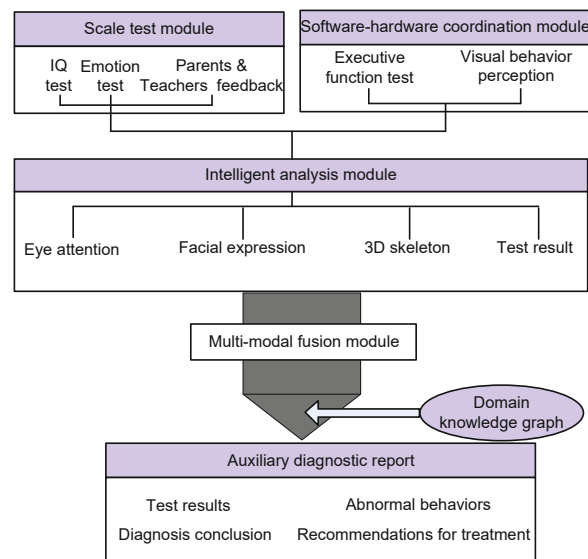


Fig. 1 Architecture of the intelligent auxiliary diagnostic system for attention deficit hyperactivity disorder (ADHD) in children

1. Scale test module

This module implements different traditional ADHD developmental psychology measures to capture and assess all aspects of children's behavioral success and related abilities. Traditional paper questionnaire approaches have issues including the complex filling process, the complicated data collection procedure, and the difficulties of further analysis. So, instead of traditional paper questionnaires, we have developed a full set of WeChat mini-programs.

Testees will fill in from their cell phones by scanning the QR code. The results of the test can be automatically integrated into the final smart diagnostic assistant report for reference by doctors.

2. Software-hardware coordination module

We take three traditional executive function tests and build correct testing software to allow the child to carry out interactive and fun-related tasks. At the same time, the hardware modules that we designed (multi-camera groups and synchronous control systems) during task testing can record attention to eye movement, facial expression, and body posture. Further smart analysis can be carried out later on the captured footage.

3. Intelligent analysis module

This module is designed to process the multimedia information obtained in the previous module, and to use computer vision technology to analyze, monitor, and classify eye movements, focus, gestures, and postures. After that, we turn them into a representation of homogenization vectors, which can be further forwarded to the deep learning model mentioned in the following session. At the same time, we document and monitor the mouse movements, clicks, keyboard input, and other behaviors of the testee, which can be incorporated for joint analysis with the above multimedia information.

4. Multi-modal fusion module

In this module, we use the time-series fusion model BERT (Devlin et al., 2019) to pre-train the similar vector obtained in the previous step, and finally generate a model that can decide if the patient has an abnormal behavior during this period. The type and frequency of an abnormal behavior serve as an essential guideline for the final report of an auxiliary diagnosis, replacing subjective findings that depend solely on the level of expertise of the doctor in conventional diagnosis.

The software-hardware coordination module, along with the following vision algorithms and deep learning model, is the core innovation point of the entire ADHD diagnostic system. The real test environment is shown in Fig. 2a. It consists of a desktop computer and camera group and is equipped with an assistant operation console to assist the test. The system has been deployed in Department of Psychology, The Children's Hospital, Zhejiang University School of Medicine and Deqing Branch, Institute of Artificial Intelligence, Zhejiang University to serve

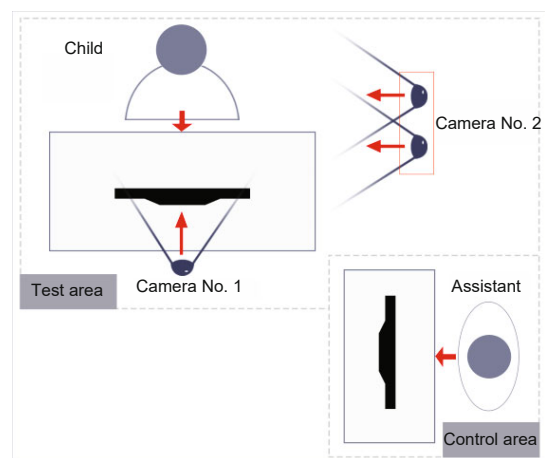
the scientific research data collection of this paper.

The schema view, as shown in Fig. 2b, describes the relationship between the cameras and the test subject. The camera No. 1 located behind the computer screen is used to capture the front image of the test subject, collecting mainly the testee's face and eye. The No. 2 binocular depth cameras located on the side of the testee's seat can be used to capture the entire body of the testee, collecting mainly the 3D body posture information of the testee. The software we developed performs executive function tests. It also provides camera module synchronization control and data acquisition, storage, and management functions. Combined with the results of the scale test, we can automatically generate a standard, objective, and quantitative interpretable auxiliary diagnostic report.

Note that this report has two main advantages: (1) The conclusions and the basis listed in our report are consistent with the existing logic of the medical diagnosis of ADHD. The key benefit lies in using



(a)



(b)

Fig. 2 Environment diagram of the software-hardware coordination module: (a) actual deployment; (b) schematic diagram

AI technology to replace the subjective findings and descriptions of doctors. (2) Combined with the current doctors' information network, we integrate the adjuvant care guidelines into the report based on the test results and diagnostic conclusions, offering higher medical value.

4 System modules

4.1 Scale test module

Table 1 lists a set of psychological scale measures widely used in ADHD diagnosis to determine the capacities of the testee, such as intellect, emotion, and social capacity. The primary purpose of the feedback scale is to carry out a preliminary assessment of the daily behavioral function of the patient. Intelligence and emotion testing is specifically to help the doctor consider the patient's condition before conducting the following functional examination, eliminating interference factors such as intellectual disability and short-term emotional disorders. That will provide a more accurate and thorough conclusion to the diagnosis.

4.2 Software-hardware coordination module

The executive function tasks are designed to evaluate multiple psychological abilities, including concentration, cognitive anti-interference, abstract thinking, cognitive conversion, emotional recognition, and social cognition. We adopt three executive function tests: Stroop test (Bench et al., 1993), Wisconsin card sorting test (WCST) (Monchi et al., 2001), and expression recognition test (Ekman, 1999; Oerlemans et al., 2014). The framework of the three tests is shown in Fig. 3. The Stroop test evaluates testees' ability to suppress cognitive interference mainly by conflicting word meanings and font colors; the WCST evaluates testees' cognitive transfer ability mainly by changing color, shape, and number of test rules; the expression recognition test primarily evaluates the social cognition ability of testees by classifying the expressions of the face pictures.

We integrate the above three executive function tasks into a set of task testing software, and the complete set of task completion time is roughly controlled within 20–30 min. In addition to functional implementation, we let professional software designers and

Table 1 Design and summary of the scale test module

Test type	Test name	Test goal
IQ test	Raven's SPM (Raven et al., 1983)	Intelligence and reasoning ability
Emotion test	SCARED (Birmaher et al., 1997)	Generalized anxiety disorder, social anxiety disorder, phobic disorders, and potential academic anxiety
	DSRSC (Birleson et al., 1987)	Depression assessment for children
Parents & Teachers feedback	CDI (Saylor et al., 1984)	Severity of depressive symptoms
	SNAP-IV (Atkins et al., 1985)	Symptoms of ADHD and ODD
	WFIRS-P (Thompson et al., 2017)	Negative mood, interpersonal problems, ineffectiveness, anhedonia, and negative self-esteem
	Conner's CBRS (Conners et al., 2011)	Academic, behavioral, and social issues

SPM: standard progressive matrices; SCARED: screen for child anxiety related disorders; DSRSC: depression self-rating scale for children; CDI: children's depression inventory; SNAP-IV: Swan-son Nolan and Pelham, version IV; ODD: oppositional defiant disorder; WFIRS-P: Weiss functional impairment rating scale-parent report; CBRS: comprehensive behavior rating scales

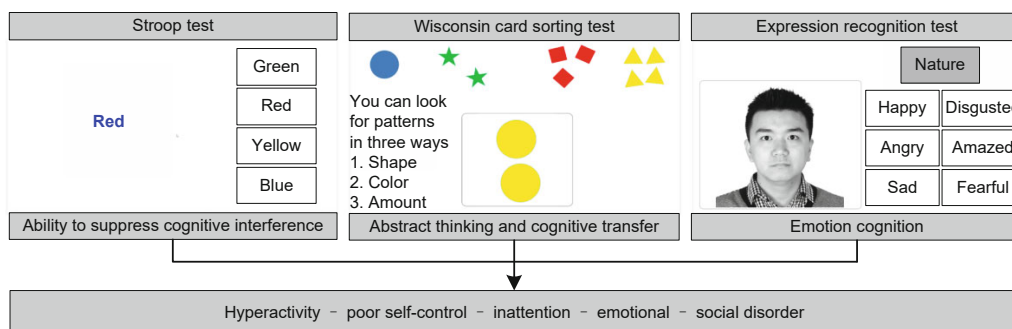


Fig. 3 Framework and purpose of the executive function tests

child psychology professionals guide software interaction and interface design, so that the tested child can complete the design task within the prescribed time.

During the three tasks, the test software can interact with the camera module. When the testee starts the test, the software will start the camera module to start recording and record the testee's eyes, expressions, and postures. The information is used for the intelligent analysis module to analyze and extract key features. When the test is completed, the test software will stop the recording of the camera module synchronously.

4.3 Intelligent analysis module

The two modules mentioned above gather and digitize the information needed. In this module, we describe how AI technology can be used to process and analyze the data collected intelligently. Our camera module collects two types of video information: (1) The camera facing the testee's face focuses on the testee's eye movement (attention focus) and facial muscle movement (abnormal expression); (2) The lateral binocular depth camera extracts the human skeleton's position at each moment using 3D visual perception technology to analyze postures. We should expand the basic methods of production of the three definitions of behavior technologies and the final vector representation of the behavior features.

4.3.1 Eye attention

The input of the eye attention model is the collected frontal video V_{front} and the camera calibration matrix C_r . The output is the generated eye attention feature vector F_g . The specific calculation process is described in Algorithm 1 and can be divided into three steps:

Step 1: Pupil position calculation

The pupil position feature for frame i is defined as a six-dimensional (6D) vector e_r , including both pupils' 3D spatial positions. First, we estimate the head's 3D location and align the landmarks on the forehead. The histogram of oriented gradient (HOG) based method detects the face position in the frame (King, 2009). When identifying multiple face regions, we will take into account the largest face bounding-box. The conceptual structure of the

Algorithm 1 Calculation of the eye attention feature

Input: Video of front camera V_{front} and front camera calibration matrix C_r

Output: Eye attention feature F_g

```

1: for  $v_i$  in  $V_{\text{front}}$  do
    // Step 1: Pupil position calculation
2: Detect facial ROI and landmarks
3: Obtain pupils' flat positions  $e_h$ 
4: Calculate head rotation matrix  $R_r$  and translation vector  $t_r$ 
5: Spatial pupil location  $e_r \leftarrow t_r + e_h$ 
    // Step 2: Gaze direction calculation
6:  $W \leftarrow C_n M C_r^{-1}$  // Transform matrix
7: Obtain  $e$  by multiplying  $W$  and  $v_i$ 
8:  $R_n \leftarrow M R_r$  // Head rotation matrix
9: Convert  $R_n$  to rotation vector  $h$ 
10: Input  $e$  and  $h$  into CNN to obtain gaze vector  $g$ 
    // Step 3: Screen position conversion
11: Calculate intersection point  $p_s$  between gaze and plane
12: Calculate region type  $r$  based on  $p_s$  and the screen structure
13: Obtain the eye attention feature for  $v_i$ ,  $f_{g_i} = [e_r, g, r]$ 
14: end for
15:  $F_g = [f_{g_1}, f_{g_2}, \dots, f_{g_N}]$ 

```

continuous conditional neural field (CCNF) is used for the identification of facial landmarks P_L (Baltrušaitis et al., 2014). Accordingly, the two pupils' flat positions, e_h , can be determined based on the positions of both eyes' corners. Then we use the efficient perspective- n -point (EPnP) algorithm to align the facial landmarks (Lepetit et al., 2009), which will align the observed face with the regular average 3D facial model F , and measure the head rotation matrix (R_r) and the translation vector (t_r) in the camera coordinate. The detected face fits into a coordinate system that constructs the eyes and mouth positions of the standard facial model F . The pose is further improved by reducing the Levenberg-Marquardt distance. The performance of that step is the spatial pupil location $e_r = t_r + e_h$.

Step 2: Gaze direction calculation

The gaze direction feature can be expressed as a 2D vector g , including two angles (yaw and pitch). To obtain the gaze vector, we need to normalize the eye image. First, we use the inverse matrix of the camera calibration matrix (C_r^{-1}) to convert the original image to a 3D position. Then we compute

the conversion matrix $M = SR$, where R is the rotation matrix and S is the scaling matrix. Fix the eye position on the z axis of the camera coordinate system and at a fixed distance from the camera. At the same time, make sure that the x axis of the head coordinate system is perpendicular to the y axis of the camera coordinate system. Thus, the normalized position is obtained. Finally, the normalized eye position is converted by a standard camera projection matrix C_n to a normalized grayscale 2D image e . In the normalized space, there is a head rotation matrix $R_n = MR_r$. To calculate the gaze vector, we transform R_n into a 2D rotation vector h . Taking h and e as inputs of a 16-layer VGGNet-based model (Simonyan and Zisserman, 2014), the output of the model is the gaze direction feature vector g .

Step 3: Screen position conversion

The screen midpoint coordinates (P_{screen}) and the screen plane average vector (F_{screen}) can be obtained from the camera coordinate system via external calibration. The line equation on which the gaze is centered can be obtained according to the angle of gaze g and the location of the eye (e_r). Then, the intersection point of the gaze line and the screen plane p_s is determined as the landing point of the sight on the screen, and the eye focus region type r is acquired according to the practical test screen structure.

In summary, the eye attention feature vector of the i^{th} frame $f_{g_i} = [e_r, g, r]$ is finally obtained by combining the output of the above steps.

4.3.2 Facial expression

The changes in expression can be defined through facial motion units and behavior of 22 types of the facial action coding system (Hamm et al., 2011). The feature is expressed as F_{exp} , where $f_{\text{exp}_i} \in F_{\text{exp}}$ is a 22D facial expression vector for frame i . Specifically, micro-expressions can be identified using a combination of regional of interests (ROI) adaptation, multi-label learning, and optimal LSTM-based temporal fusion structure. The input of the measurement of the expression feature is the collected frontal video V_{front} , and the output is the F_{exp} function sequence created by the facial action unit (AU). The calculation process is outlined in Algorithm 2. Based on the physiological structure of the face, the location of the corresponding muscle linked to the expression motor unit

P_{AU} can then be determined through the fixed mode conversion, depending on the position of these landmarks P_L from Section 4.3.1. Depending on this, we can create ROI cropping networks (ROI Nets), depending on VGGNet (Simonyan and Zisserman, 2014). The function representation corresponding to each AU is obtained by cutting out the network's 12th layer feature diagram. Around that moment, all the function vectors are concatenated to obtain the overall F_{AU} vector describing the expression features.

Because the input form of the expression recognition task is video data, we can more reliably and smoothly predict the state of expression of the current moment based on the state of expression of the previous moments. Therefore, we use a multi-layer LSTM structure (Graves and Schmidhuber, 2005) to process the feature vectors for the time-series expression. The multi-task binary classification problem for multiple AUs is achieved by comparing the expression features at the current moment with the backward state, and the facial expression feature sequence is generated based on the AU feature activation probability called F_{exp} .

Algorithm 2 Calculation of the facial expression feature

Input: Video of front camera V_{front}

Output: Facial expression feature F_{exp}

- 1: **for** v_i in V_{front} **do**
 - 2: Detect facial ROI
 - 3: Detect facial landmarks P_{L_i}
 - 4: Derive muscle position P_{AU_i} from P_{L_i} according to the fixed method
 - 5: Use ROI Nets to obtain the feature vector of AU, F_{AU_i}
 - 6: **end for**
 - 7: Obtain facial feature sequence F_{exp} from F_{AU} by multi-task binary classification of LSTM
-

4.3.3 3D skeleton

The input of the 3D skeleton calculation is the captured side video V_{side} , and the output is the generated 3D pose feature sequence F_{3D} . For frame i , the motion feature is expressed as f_{3D_i} ($f_{3D_i} \in F_{3D}$), including 3D coordinates for 25 key points. The specific calculation process is described in Algorithm 3, and it consists mainly of three steps:

Step 1: Confidence map and PAF calculation

For a frame of the video, we extract the image features \mathbf{F} through the 10th layer of VGG-19 (Simonyan and Zisserman, 2014), and then input \mathbf{F} into two multi-stage CNNs. The first network produces detection confidence maps to obtain the position of critical points, and the second network determines the limb orientation through part affinity fields (PAF) (Cao et al., 2017). In the first stage, the inputs of both CNNs are image features \mathbf{F} , and the confidence maps $\mathbf{S}^1 = \rho^1(\mathbf{F})$ and PAF $\mathbf{L}^1 = \phi^1(\mathbf{F})$ are the output. Afterward, the confidence map and PAF of the t^{th} subsequent stage can be expressed as

$$\begin{cases} \mathbf{S}^t = \rho^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2, \\ \mathbf{L}^t = \phi^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2. \end{cases} \quad (1)$$

We can further improve the accuracy through the results of multiple stages, and obtain the final output containing the confidence maps $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_J)$ and PAF $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_C)$, where J and C represent the type number of key points and limbs respectively.

Algorithm 3 Calculation of the 3D skeleton feature

Input: Video of side camera \mathbf{V}_{side}

Output: 3D posture feature \mathbf{F}_{3D}

```

1: for  $\mathbf{v}_i$  in  $\mathbf{V}_{\text{side}}$  do
    // Step 1: Confidence map and PAF calculation
2: Extract image features  $\mathbf{F}$  from  $\mathbf{v}_i$ 
3: for Stage  $t$  in CNN do
4:   if  $t=1$  then
5:      $\mathbf{S}^1 \leftarrow \rho^1(\mathbf{F})$  // Maps
6:      $\mathbf{L}^1 \leftarrow \phi^1(\mathbf{F})$  // PAF
7:   else
8:      $\mathbf{S}^t = \rho^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1})$ 
9:      $\mathbf{L}^t = \phi^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1})$ 
10:  end if
11: end for
    // Step 2: 2D posture generation
12: for  $c \leftarrow 1$  to  $C$  do
13:   Optimize weight  $E_c$  by the Hungarian algorithm
14: end for
15: Connect all limbs with common key points to
    generate 2D pose sequence  $\mathbf{f}_{2D_i}$ 
    // Step 3: 3D posture conversion
16: Convert  $\mathbf{f}_{2D_i}$  to  $\mathbf{f}_{3D_i}$  by a neural network
17: end for
18:  $\mathbf{F}_{3D} = [\mathbf{f}_{3D_1}, \mathbf{f}_{3D_2}, \dots, \mathbf{f}_{3D_N}]$ 

```

Step 2: 2D posture generation

According to the confidence map, we can obtain the positions of key points. Letting $\mathbf{d}_{j_1}^m$ indicate the

m^{th} detected point which belongs to the j^{th} type, then for the key point pair $(\mathbf{d}_{j_1}^m, \mathbf{d}_{j_2}^n)$ about limb type c , we can calculate the confidence $E_{j_1 j_2}^{mn}$ whether the points connect with \mathbf{L}_c :

$$E_{j_1 j_2}^{mn} = \int_{u=0}^{u=1} L_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2}^n - \mathbf{d}_{j_1}^m}{\|\mathbf{d}_{j_2}^n - \mathbf{d}_{j_1}^m\|} du, \quad (2)$$

where \mathbf{p} is the interpolation function between the two points:

$$\mathbf{p}(u) = (1-u)\mathbf{d}_{j_1}^m + u\mathbf{d}_{j_2}^n. \quad (3)$$

The zero-one variable $z_{j_1 j_2}^{mn}$ indicates whether $\mathbf{d}_{j_1}^m$ is connected to $\mathbf{d}_{j_2}^n$. Our task shifts to maximizing the weight of selected connections:

$$\begin{aligned} \max_{\mathcal{Z}_c} E_c &= \max_{\mathcal{Z}_c} \sum_{m \in \mathcal{D}_{j_1}} \sum_{n \in \mathcal{D}_{j_2}} E_{j_1 j_2}^{mn} \cdot z_{j_1 j_2}^{mn} \\ \text{s.t.} \quad &\begin{cases} \forall m \in \mathcal{D}_{j_1}, \sum_{n \in \mathcal{D}_{j_2}} z_{j_1 j_2}^{mn} \leq 1, \\ \forall n \in \mathcal{D}_{j_2}, \sum_{m \in \mathcal{D}_{j_1}} z_{j_1 j_2}^{mn} \leq 1, \end{cases} \end{aligned} \quad (4)$$

where E_c represents the total weight of limb c in the current matching method, \mathcal{Z}_c represents the set of z about limb c , and \mathcal{D}_j represents the set of key points for the j^{th} type.

To reduce the computation, we prune the full connection, keep only the connections between adjacent points, and decompose the problem into a set of pairwise matching problems between key points. The overall optimization goal can be expressed as

$$\max_{\mathcal{Z}} E = \sum_{c=1}^C \max_{\mathcal{Z}_c} E_c. \quad (5)$$

We can use the Hungarian algorithm to optimize such a problem (Kuhn, 1955), and connect the limbs with common key points to form a complete human 2D pose feature \mathbf{f}_{2D_i} .

Step 3: 3D posture conversion

Finally, we input the 2D posture feature \mathbf{f}_{2D_i} into a neural network to obtain the final output 3D pose sequence \mathbf{f}_{3D_i} (Martinez et al., 2017). Thus, for the whole video, the 3D skeleton feature $\mathbf{F}_{3D} = [\mathbf{f}_{3D_1}, \mathbf{f}_{3D_2}, \dots, \mathbf{f}_{3D_N}]$ is observed.

5 Abnormal behavior detection

The core task of diagnosing ADHD is to observe the presence of abnormal behaviors. From the video recordings, we extract three separate dimensions of

eye focus, speech, and posture in the intelligent analysis module. Via in-depth consultation with qualified physicians, we outline many examples of significant irregular behaviors often exhibited by children with ADHD during functional testing (Table 2). The key idea is to replace the doctor's observations of children's behaviors by using AI technology to automatically detect fragments of abnormal behavior in the videos. More precisely, we are proposing a model based on BERT for a multi-modal fusion of knowledge. The performance of the smart research module is combined with the functional test results to train an abnormal behavioral classification model. The detection algorithm is divided primarily into the following two steps:

1. Video segmentation

As seen in Fig. 4, the entire testee video takes about 20–30 min. We can divide the entire video into several segments based on the testees' completion situation of the executive function tests.

2. Segment classification

We combine the performance of the intelligent analysis module with the functional test result to form the instant feature vector for each section. Moreover, we concatenate the instantaneous vectors to form time-series features representing time segments that can be used to evaluate if the testee has irregular behaviors.

Table 2 Description of abnormal behavior types

Behavior type	Specific behaviors
Eye attention	Erratic eyes
	Incorrect focus position
Facial expression	Facial muscle twitching
	Frequent head shaking
	Hand-to-mouth movement
Body posture	Leg shaking
	Twitching
	Playing with microphone or cameras

5.1 BERT

BERT is a landmark model in the NLP field (Devlin et al., 2019). It once smashed the record of 11 tasks in the NLP sector, when it was proposed in 2018. Great for handling sequence problems, BERT analyzes the relationship between items in the whole sequence of inputs to encode the current item. The BERT model structure is shown in Fig. 5. It is composed of 12 transformer encoder layers. Each layer of encoders comprises a layer of attention and forward feedback (Vaswani et al., 2017). In the attention layer, the point product attention function for the input x is determined to obtain a vector group Z , which represents the impact weight of each sequence element on the current item:

$$Z = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (6)$$

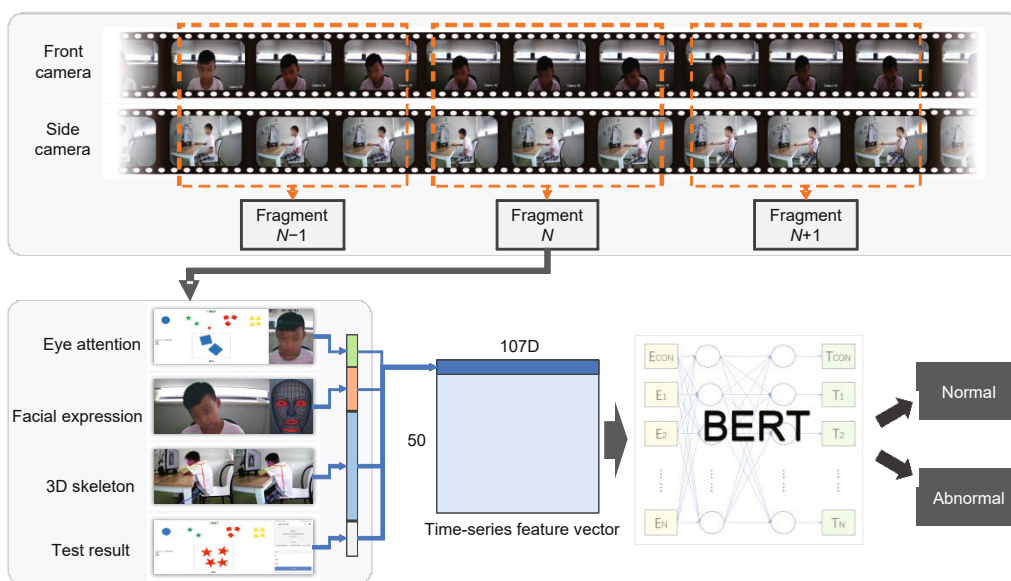


Fig. 4 Framework of the abnormal behavior detection algorithm

where Q , K , and V are hidden variables in the attention mechanism, and $\sqrt{d_k}$ is the square root of the vector dimension of vector group K . For the multi-head attention model, it is necessary to use multiple independent sets of Q , K , V hidden variables to generate multiple vector groups $\{Z_1, Z_2, \dots, Z_T\}$, where T is the total number of vector groups to obtain a better model expression ability. The weighted summation of all vector groups, Z_{final} , can be explained as a multi-head attention result:

$$Z_{\text{final}} = \text{Concat}(Z_1, Z_2, \dots, Z_T)W^o, \quad (7)$$

where W^o is a dimension-reduction matrix.

The forward feedback layer consists of two fully connected layers that are responsible for processing the attention layer output further. Additionally, residuals are applied to both the attention layer and the forward feedback layer to prevent gradients from disappearance during back propagation when the number of encoder layers is large.

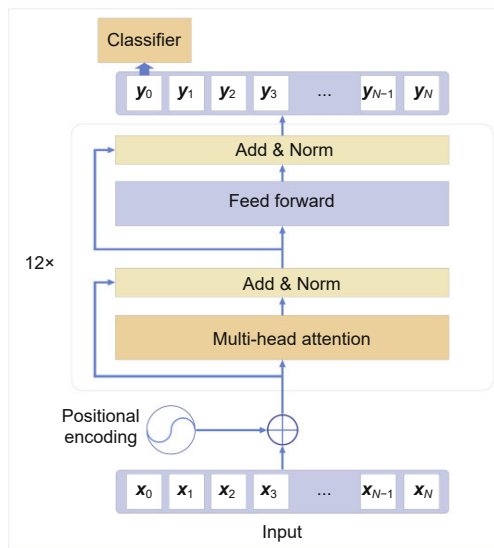


Fig. 5 Framework of the BERT block

Back to our problem of identifying abnormal behaviors, the behavioral vectors created by the intelligent analysis module can be used as sequence input. To make a final decision, the analysis of the feature vector v_i must be combined each time with the feature vectors of other moments in the behavior feature vector group V . In the following subsections, we will elaborate on the algorithm and the accuracy of detection.

5.2 Detection algorithm

For the entire answering process, there are front video V_{front} , side video V_{side} , and answer operation sequence $A = \{a_1, a_2, \dots, a_K\}$, where a_i ($i = 1, 2, \dots, K$) represents the timestamp of answering the i^{th} question and K is the total number of answers. As the estimated answer time is 2 s, K' periods can be selected from the entire answering process. For each time period, in the case where the frame rate is 25 frames/s, the video frames corresponding to V_{front} and V_{side} can form an image sequence in length 50. The corresponding visual feature is extracted from the video frames, including eye attention vector F_g , facial expression vector F_{exp} , and posture vector F_{3D} .

For moment i , we concatenate three different vectors (f_{g_i} (9D), f_{exp_i} (22D), and f_{3D_i} (75D)) obtained by the intelligent analysis module and the current answer f_{ans_i} (which forms the final vector representing the behavior characteristics of the tettee at that time). For a fragment, we can obtain the behavior feature $x_i = [f_{g_i}, f_{\text{exp}_i}, f_{3D_i}, f_{\text{ans}_i}]$ (107D) representing the current moment. Therefore, for the entire test, we can obtain the 107D behavior feature vector group $X = \{x_0, x_1, x_2, \dots, x_N\}$, where x_0 is a constant vector.

The output vector group obtained by BERT is recorded as $Y = \{y_0, y_1, y_2, \dots, y_N\}$, where y_0 is the output corresponding to x_0 , which represents the vector representation of the entire time segment. To send y_0 to a fully connected layer classifier, we can determine whether this fragment contains abnormal behaviors. Considering that no matter for the ADHD patients or healthy children, the proportion of normal fragments occupies the majority, we use weighted cross-entropy as the loss function to train the model to alleviate the problem of data imbalance.

5.3 Experiment

We collected diagnostic data from 82 children at The Children's Hospital, Zhejiang University School of Medicine from August 2019 to December 2019. The data for each child included 7 scale test results, 3 executive function test results, and 2 camera group video shots of the patient's face and side body. We carried out manual marking of irregular behaviors on the 53-h videos obtained, under the supervision of the medical team.

5.3.1 Labeling rules

According to the video duration and frame rate, if more than 30% of the time segment is marked as an abnormal activity, the segment is considered to be an abnormal segment; otherwise, it is considered to be a normal one. We can divide the entire video into a large number of shorter time segments, thus obtaining enough time fragments to support an abnormal behavior detection model for deep learning.

5.3.2 Dataset

The 82 children we recruited include 71 patients with ADHD and 11 non-patients (Table 3). We received 9116 abnormal fragments from 71 patients with ADHD after the manual marking mentioned above, which were further divided into two subsets (6700 for training and 2416 for testing). Similarly, from 11 healthy children, we obtained 27 666 normal fragments, which were further divided into two subsets (24 461 for training and 3205 for testing). Note that only a small portion of fragments were abnormal for the children with ADHD, so even if the abnormal children were with the majority, the number of abnormal fragments was not significant. Implementation details were as follows: The batch normalization scale was set at 5 during the training phase, the model's initial learning rate was set at 1×10^{-4} , and the cosine annealing scheduler was used as the learning rate adjustment technique. To optimize the model, we used the stochastic gradient descent (SGD) optimization algorithm and each model was trained for 100 epochs.

5.3.3 Results

The test results of the detection module for abnormal behaviors are shown in Table 4. We measured the model's performance through different indicators including accuracy, sensitivity, specificity, false positive rate (FPR), and false negative rate (FNR). Also, we have added two conventional deep

sequential models (GRU and LSTM) as control experiments. From the tests, it can be seen that our model achieved 80.25% accuracy in detection. Our model maintained a high sensitivity (0.9357), keeping the excellent performance of avoiding misjudging the normal behaviors into abnormal. Considering that the vast majority of normal fragments were occupied, the higher sensitivity ensured that the results were valid. On this basis, our model obtained a good specificity result (0.6258).

5.3.4 Multi-modal priority

To determine the importance of specific multi-modal data for abnormal behavior identification, Table 4 also offers four ablation experiments, separately masking eye focus information (BERT-eye), facial expression (BERT-fac), 3D-skeleton (BERT-ske), and test response (BERT-ans). From the results, we can see that the effect of 3D-skeleton on detecting abnormal behaviors was the most significant, with minimum impact on eye attention. We believe that this finding was triggered by the following two principal reasons: (1) Hyperactivate is the most common, direct, and significant symptom of ADHD, whereas eye attention abnormality is relatively vague and challenging to observe; (2) The action contains high-dimensional and adequate information, while the eye movement features are relatively small in dimension. We are noting that the response situation has a significant influence, showing that through this kind of knowledge, we can differentiate abnormal behaviors from normal large-scale information.

6 Intelligent system

6.1 Auxiliary diagnostic system

We have developed a complete ADHD diagnostic process system, including a set of scale tests, execution function tests, an AI algorithm suite, and a platform for information management.

1. Scale test tool

Table 3 Description of the video fragment dataset

Testee group	Training set	Testing set	Total
Number of patients	58	13	71
Number of abnormal fragments	6700	2416	9116
Number of non-patients	9	2	11
Number of normal fragments	24 461	3205	27 666

Table 4 Results of abnormal behavior detection

Method	Accuracy (%)	Sensitivity	Specificity	FPR	FNR
GRU	73.10	0.8720	0.5439	0.4561	0.1280
LSTM	76.26	0.8933	0.5894	0.4106	0.1067
BERT (ours)	80.25	0.9357	0.6258	0.3742	0.0643
BERT-eye	76.65	0.9045	0.5836	0.4164	0.0955
BERT-fac	74.95	0.8689	0.5911	0.4089	0.1311
BERT-ske	66.09	0.8970	0.3477	0.6523	0.1030
BERT-ans	74.26	0.8633	0.5837	0.4163	0.1367

FPR: false positive rate; FNR: false negative rate. The best performance of each metric is in bold

Patients can register personal information and perform various scale tests via WeChat, based on the WeChat mini-program platform.

2. Execution function test software

On the PC side, the software is implemented and created through C # WPF. The software is to simultaneously accomplish the executive function testing and camera module control and report the user operations and timestamps.

3. AI algorithm suite

The suite is developed using the languages C++ and Python. It is implemented using the deep learning framework of Pytorch 1.0.1 and uses the general parallel computing architecture of CUDA 10.0 to implement GPU operations. For the processing of images, we use the openCV library.

4. Information management platform

We have created a web-app that is deployed on Alibaba Cloud. It uses the SQL Server database to implement the storage and management of information. The core role of this platform is to produce the patient's auxiliary diagnostic report.

6.2 Auxiliary diagnostic report

We can automatically generate an auxiliary diagnostic report after completing the entire ADHD testing process (Fig. 6). The report includes scale tests (IQ, emotion, social functioning, etc.), executive function test results (Stroop test, WCST, comprehension of expression), and the nature and frequency of repetitive behaviors observed in recorded videos. At the same time, diagnostic conclusions and treatment guidelines will be automatically generated based on the ADHD medical information network as well as the detailed test results. In the right-top corner of the diagnostic report, doctors and patients can scan the QR code to obtain the detailed results for each individual test.

Diagnostic Report

Name: Zhang San Age: 8

Gender: Male Date: 2019/10/26

Please scan the wechat code for complete test results

Triage Tests

SNAP-IV (Attention Rating Scale)
Test Result (Parents):
 Attention Deflect: Severe,
 Hyperactive Impulse: Moderate,
 Oppositional Defiance: Normal
Test Result (Teachers):
 Attention Deflect: Severe,
 Hyperactive Impulse: Moderate,
 Oppositional Defiance: Normal

WEISS (Social Ability Scale)
Defective Section (Parents): Study and School, Life Scale, Social Activity.
Defective Section (Teachers): Study and School, Self Concept, Social Activity, Adventure Activity, Cognition and Emotion.

Conners Questionnaire for Parents
Abnormal Factors: Learning Problem.
 IQ / Emotional Tests

Raven's SPM
 IQ: 79, Percent Intelligence: 21%, Quiz Time: 08:45

SCARED (Screen for Child Anxiety Related Disorders)
 Score 26, above the demarcation value(23). Abnormal
 Abnormal factor: Separation anxiety, Social horror

DSRSC (Depression Self-Rating Scale for Children)
 Score 16, above the demarcation value(23). Abnormal

CDI (Children's Depression Inventory)
 Score 13, below the demarcation value(19). Normal
 Task / Behavior test

Stroop Test
 No obvious abnormalities

WCST (Wisconsin Card Sorting Test)
 Abnormal cognitive transfer ability, poor abstract summary ability, especially poor initial concept formation ability, poor insight in concept formation, poor cognitive transfer ability, inattention or confused thinking

Expression Test
 No obvious abnormalities

Visual assessment
Stroop Test:
 04:15 - 04:36 abnormal behavior
 05:20 - 05:35 abnormal behavior
 05:57 - 06:05 abnormal behavior
Wisconsin Card Sorting Test:
 07:55 - 08:02 abnormal behavior
 08:20 - 08:37 abnormal behavior
 09:42 - 09:53 abnormal behavior
 10:16 - 10:28 abnormal behavior
Expression Test:
 12:37 - 13:02 abnormal behavior
 14:12 - 14:20 abnormal behavior
 16:47 - 16:54 abnormal behavior
 17:33 - 17:59 abnormal behavior

Assessment result:
 The child is not attentive, and has poor self-control
 Diagnosis conclusion / Treatment opinion

Diagnosis conclusion
 Inattentive ADHD

Treatment opinion
Behavioral therapy: Cognitive activities appropriate to the patient's characteristics should be used to improve their attention.
Self-control training: Through simple and fixed self-training commands, the patient will learn to control themselves.
Others: Let the patient eat more healthy foods and limit the foods with formaldehyde and salicylic acid.

Fig. 6 Real case of an ADHD diagnostic report

7 Conclusions

The conventional medical diagnosis of ADHD is frequently restricted by the lack of medical resources, resulting in insufficient monitoring of children at the clinic. This is also highly reliant on the level of expertise of the doctor. Hence, we have designed an

auxiliary diagnostic system to fully leverage the AI technology. During the completion of related functional tests, our specially designed software and hardware system can record children's behaviors and intelligently analyze the eye attention, facial expressions, and information on body postures. We have proposed a deep learning model (BERT) to fuse all information for the identification of abnormally behaved segments, with all the visual signals needed. The entire information system standardizes the method of ADHD diagnosis. It provides detailed assessment reports, including scale test findings, findings of executive function evaluations, abnormal behavior review, implications of medical aids, and treatment recommendations.

Our system has been carrying out auxiliary diagnostic work for several months at The Children's Hospital, Zhejiang University School of Medicine, which has greatly improved doctors' efficiency and provided children with a rich and standardized diagnostic report. First, just an operative assistant with no medical training is needed, and the patient will perform the executive function tests of 20–30 min properly for a more abundant evaluation. The application of intelligent analysis complemented the insufficient assessment of patients by doctors, which decreases the time cost of outpatient assessment. Thereby, the average consultation time is reduced from 15–20 min to around 10 min, increasing the turnaround performance of outpatient. At the same time, comprehensive and standardized diagnostic reports are generated automatically and allow doctors to obtain a faster diagnosis, reducing the potential for inconclusive diagnosis, misdiagnosis, and the unnecessary follow-up treatment. Adequate diagnostic foundation decreases the risk for difficult diagnosis and misdiagnosis and thereby induces unnecessary follow-up diagnosis. Additionally, although it takes the AI system about 30 min to analyze the video, patients do not have to wait for hospital outcomes. They will subsequently receive more detailed and standardized electronic diagnostic reports, with no harm to the treatment experience.

We plan to promote this system to more hospitals in the future, shortly, to gather more clinical data. We will constantly track the recovery process of the patients using the system and quantitatively analyze the clinical diagnosis support from our program. We will also develop the algorithms

and the auxiliary program, and take the diagnosis of more related diseases and complications into account so that we can empower more patients with the AI technology.

Contributors

Yanyi ZHANG, Ming KONG, Wenchen HONG, and Qiang ZHU designed the research. Yanyi ZHANG, Rongwang YANG, and Rong LI provided the test environment and helped collect the data. Wenchen HONG and Tianqi ZHAO processed the data. Ming KONG, Wenchen HONG, Tianqi ZHAO, Di XIE, and Chunmao WANG designed and developed the system. Ming KONG and Tianqi ZHAO drafted the manuscript. Yanyi ZHANG and Qiang ZHU helped organize the manuscript. Ming KONG, Tianqi ZHAO, and Qiang ZHU revised and finalized the paper.

Compliance with ethics guidelines

Yanyi ZHANG, Ming KONG, Tianqi ZHAO, Wenchen HONG, Di XIE, Chunmao WANG, Rongwang YANG, Rong LI, and Qiang ZHU declare that they have no conflict of interest.

The study protocol was approved by the Medical Ethics Committee in The Children's Hospital, Zhejiang University School of Medicine (2018-IRB-003) and was conducted in accordance with the Helsinki Declaration of 1975, as revised in 2008 (5). Informed consent was obtained from all patients for being included in the study.

References

- Aradhya AMS, Joglekar A, Suresh S, et al., 2019. Deep transformation method for discriminant analysis of multi-channel resting state fMRI. *Proc AAAI Conf on Artificial Intelligence*, p.2556-2563. <https://doi.org/10.1609/aaai.v33i01.33012556>
- Atkins MS, Pelham WE, Licht MH, 1985. A comparison of objective classroom measures and teacher ratings of attention deficit disorder. *J Abnorm Child Psychol*, 13(1):155-167. <https://doi.org/10.1007/BF00918379>
- Baltrušaitis T, Robinson P, Morency LP, 2014. Continuous conditional neural fields for structured regression. *Proc 13th European Conf on Computer Vision*, p.593-608. https://doi.org/10.1007/978-3-319-10593-2_39
- Bench CJ, Frith CD, Grasby PM, et al., 1993. Investigations of the functional anatomy of attention using the Stroop test. *Neuropsychologia*, 31(9):907-922. [https://doi.org/10.1016/0028-3932\(93\)90147-R](https://doi.org/10.1016/0028-3932(93)90147-R)
- Birleson P, Hudson I, Buchanan DG, et al., 1987. Clinical evaluation of a self-rating scale for depressive disorder in childhood (depression self-rating scale). *J Child Psychol Psych*, 28(1):43-60. <https://doi.org/10.1111/j.1469-7610.1987.tb00651.x>

- Birmaher B, Khetarpal S, Brent D, et al., 1997. The screen for child anxiety related emotional disorders (SCARED): scale construction and psychometric characteristics. *J Am Acad Child Adolesc Psych*, 36(4):545-553.
<https://doi.org/10.1097/00004583-199704000-00018>
- Cao Z, Simon T, Wei SE, et al., 2017. Realtime multi-person 2D pose estimation using part affinity fields. *Proc IEEE Conf on Computer Vision and Pattern Recognition*, p.7291-7299.
<https://doi.org/10.1109/CVPR.2017.143>
- Chen M, Li HL, Wang JH, et al., 2019. A multichannel deep neural network model analyzing multiscale functional brain connectome data for attention deficit hyperactivity disorder detection. *Radiol Artif Intell*, 2(1):e190012.
<https://doi.org/10.1148/ryai.2019190012>
- Conners CK, Pitkanen J, Rzepa SR, 2011. Conners comprehensive behavior rating scale. In: Kreutzer JS, DeLuca J, Caplan B (Eds.), *Encyclopedia of Clinical Neuropsychology*. Springer, New York, USA, p.678-680.
https://doi.org/10.1007/978-0-387-79948-3_1536
- Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p.4171-4186.
<https://doi.org/10.18653/v1/N19-1423>
- Ekman P, 1999. Basic emotions. In: Dalglish T, Dalglish MJ (Eds.), *Handbook of Cognition and Emotion*. Wiley, New York, USA, p.301-320.
<https://doi.org/10.1002/0470013494.ch3>
- Graves A, Schmidhuber J, 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neur Netw*, 18(5-6):602-610.
<https://doi.org/10.1016/j.neunet.2005.06.042>
- Hamm J, Kohler CG, Gur RC, et al., 2011. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *J Neurosci Meth*, 200(2):237-256.
<https://doi.org/10.1016/j.jneumeth.2011.06.023>
- Jaiswal S, Valstar MF, Gillott A, et al., 2017. Automatic detection of ADHD and ASD from expressive behaviour in RGBD data. *Proc 12th IEEE Int Conf on Automatic Face & Gesture Recognition*, p.762-769.
<https://doi.org/10.1109/FG.2017.95>
- King DE, 2009. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*, 10:1755-1758.
- Kuhn HW, 1955. The Hungarian method for the assignment problem. *Nav Res Logist Q*, 2(1-2):83-97.
<https://doi.org/10.1002/nav.3800020109>
- Leo M, Carcagni P, Distanti C, et al., 2018. Computational assessment of facial expression production in ASD children. *Sensors*, 18(11):3993.
<https://doi.org/10.3390/s18113993>
- Lepetit V, Moreno-Noguer F, Fua P, 2009. EPnP: an accurate $O(n)$ solution to the PnP problem. *Int J Comput Vis*, 81(2):155.
<https://doi.org/10.1007/s11263-008-0152-6>
- Li J, Zhong YH, Han JX, et al., 2019. Classifying ASD children with LSTM based on raw videos. *Neurocomputing*, 390:226-238.
<https://doi.org/10.1016/j.neucom.2019.05.106>
- Marcano JL, Bell MA, Beex AAL, 2018. Classification of ADHD and non-ADHD subjects using a universal background model. *Biomed Signal Process Contr*, 39:204-212.
<https://doi.org/10.1016/j.bspc.2017.07.023>
- Martinez J, Hossain R, Romero J, et al., 2017. A simple yet effective baseline for 3d human pose estimation. *Proc IEEE Int Conf on Computer Vision*, p.2640-2649.
<https://doi.org/10.1109/ICCV.2017.288>
- Monchi O, Petrides M, Petre V, et al., 2001. Wisconsin card sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *J Neurosci*, 21(19):7733-7741.
<https://doi.org/10.1523/JNEUROSCI.21-19-07733.2001>
- Muñoz-Organero M, Powell L, Heller B, et al., 2019. Using recurrent neural networks to compare movement patterns in ADHD and normally developing children based on acceleration signals from the wrist and ankle. *Sensors*, 19(13):2935. <https://doi.org/10.3390/s19132935>
- Oerlemans AM, van der Meer JM, van Steijn DJ, et al., 2014. Recognition of facial emotion and affective prosody in children with ASD (+ADHD) and their unaffected siblings. *Eur Child Adolesc Psych*, 23(5):257-271.
<https://doi.org/10.1007/s00787-013-0446-2>
- Polaczyk GV, Willcutt EG, Salum GA, et al., 2014. ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int J Epidemiol*, 43(2):434-442.
<https://doi.org/10.1093/ije/dyt261>
- Raven JC, Raven JH, 1983. *Manual for Raven's Progressive Matrices and Vocabulary Scales: Standard Progressive Matrices*. Lewis, London, UK.
- Sayal K, Prasad V, Daley D, et al., 2018. ADHD in children and young people: prevalence, care pathways, and service provision. *Lancet Psych*, 5(2):175-186.
[https://doi.org/10.1016/S2215-0366\(17\)30167-0](https://doi.org/10.1016/S2215-0366(17)30167-0)
- Saylor CF, Finch AJ, Spirito A, et al., 1984. The children's depression inventory: a systematic evaluation of psychometric properties. *J Consult Clin Psychol*, 52(6):955-967.
<https://doi.org/10.1037/0022-006X.52.6.955>
- Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition.
<https://arxiv.org/abs/1409.1556>
- Thompson T, Lloyd A, Joseph A, et al., 2017. The Weiss functional impairment rating scale-parent form for assessing ADHD: evaluating diagnostic accuracy and determining optimal thresholds using ROC analysis. *Qual*

- Life Res*, 26(7):1879-1885.
<https://doi.org/10.1007/s11136-017-1514-8>
- Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.5998-6008.
- Wang TT, Liu KH, Li ZZ, et al., 2017. Prevalence of attention deficit/hyperactivity disorder among children and adolescents in China: a systematic review and meta-analysis. *BMC Psych*, 17(1):32.
<https://doi.org/10.1186/s12888-016-1187-9>
- Willcutt EG, Nigg JT, Pennington BF, et al., 2012. Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *J Abnorm Psychol*, 121(4):991-1010.
<https://doi.org/10.1037/a0027347>
- Zou L, Zheng JN, Miao CY, et al., 2017. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. *IEEE Access*, 5:23626-23636.
<https://doi.org/10.1109/ACCESS.2017.2762703>