sciendo

Int. J. Appl. Math. Comput. Sci., 2020, Vol. 30, No. 3, 435–448
DOI: 10.34768/amcs-2020-0032

amcs

AN INTELLIGENT MULTIMODAL FRAMEWORK FOR IDENTIFYING CHILDREN WITH AUTISM SPECTRUM DISORDER

JINGYING CHEN [a,b], MENGYI LIAO [a,c,*], GUANGSHUAI WANG [a], CHANG CHEN [a]

[a]National Engineering Research Center for E-Learning
Central China Normal University
Wuhan 430079, Hubei, China
e-mail: liaomengyi83@163.com

[b]National Engineering Laboratory for Technology of Big Data Application in Education
Central China Normal University
Wuhan 430079, Hubei, China

[c]College of Computer Science and Technology
Pingdingshan University
Pingdingshan 467000, Henan, China

Early identification can significantly improve the prognosis of children with autism spectrum disorder (ASD). Yet existing identification methods are costly, time consuming, and dependent on the manual judgment of specialists. In this study, we present a multimodal framework that fuses data on a child's eye fixation, facial expression, and cognitive level to automatically identify children with ASD, to improve the identification efficiency and reduce costs. The proposed methodology uses an optimized random forest (RF) algorithm to improve classification accuracy and then applies a hybrid fusion method based on the data source and time synchronization to ensure the reliability of the classification results. The classification accuracy of the framework was 91%, which is higher than that of the RF, support vector machine, and discriminant analysis methods. The results suggest that data on a child's eye fixation, facial expression, and cognitive level are useful for identifying children with ASD. Because the proposed framework can separate ASD children from typically developing (TD) children, it can facilitate the early identification of ASD and may improve intervention programs for children with ASD.

Keywords: autism spectrum disorder, eye fixation, facial expression, cognitive level, improved random forest.

1. Introduction

Autism spectrum disorder (ASD) is a broad neurodevelopmental disorder characterized by social communication disorders, verbal and non-verbal communication deficits, narrow interests, and repetitive and rigid behaviors (Amaral *et al*., 2008). Currently, the cause of ASD is not clear, and there are no drugs that can cure it. Because most people with ASD have social maladjustments or lifelong disorders, they cannot take care of themselves, which places an economic and mental burden on society and their families. The World Health Organization (WHO) identifies ASD as a growing problem that seriously impacts the quality of life (Durkin *et al*., 2010). Recently, the number of children diagnosed with ASD has increased dramatically. In terms of global prevalence, roughly 1 in 160 children suffer from ASD. In America, 1 in 59 children is diagnosed with ASD according to the Autism and Developmental Disabilities Monitoring (ADDM) Network (Christensen *et al*., 2016).

Early identification can dramatically improve the prognosis of children with ASD, as the rapid brain development is beneficial to treatment. Unfortunately, there are no efficient tools to automate early identification, and most medical institutes do not have enough ASD specialists (Zwaigenbaum *et al*., 2009). According to statistics, ASD identification is usually delayed until the age of 4, and approximately 27% of cases are delayed until the age of 8 (Halim *et al*., 2018). Most identification tools

*Corresponding author

in use today are based on standardized questionnaires for parents or specialists, such as the Modified Checklist for Autism in Toddlers (M-CHAT) (Bernier *et al*., 2011) and the Child Behavior Checklist (CBCL) (Achenbach and Rescorla, 2000). Specialists administer identification tools in rigorously controlled clinical settings, and they usually take several hours to complete (Jaiswal *et al*., 2017). Because these manual tools are time consuming and difficult to apply, there are major technological obstacles to identifying ASD in children.

Studies on children with ASD have generated an enormous amount of treatment and diagnostic data. Following the development of information and communication technologies, such as mobile Internet, smart sensors, and cloud computing, data-driven intelligent analytics are increasingly used in the fields of medicine and education (Xu *et al*., 2017). For example, a medical study may analyze the characteristic data of its patients to determine the best treatment approach, with the added benefits of both improving treatment efficiency and reducing medical costs. Currently, a lot of data for ASD diagnosis are being generated. Historical data should be used during diagnosis and treatment as a basis of judgment for isolate children with ASD.

Machine learning and smart sensor data have recently been used for the early identification of autism, resulting in a simple, low-cost method that holds promise in distinguishing between ASD children and typically developing (TD) children. Tariq *et al*. (2018) utilized a machine learning algorithm to analyze home videos of children, which greatly sped up the diagnosis of ASD in those cases with accurate outcomes. Using a recurrent deep neural network, Zunino *et al*. (2018) similarly processed video clips of children grasping a bottle and accurately distinguished the children with ASD from TD. Jiang *et al*. (2019) also proposed a machine learning method based on eye fixation to differentiate ASD from TD, achieving a classification accuracy of 86%. However, although these identification methods show promise, most focused on a single-modal approach, which is not diversified enough.

In fact, researchers have proposed a variety of theoretical models to explain the behavioral or cognitive abnormalities of individuals with ASD from different perspectives, which can be used as the theoretical basis for the early identification of ASD. From the perspective of social cognitive, there are the theory of mind and the theory of weak central coherence. From the perspective of neuropsychology, there is the theory of broken mirror. According to the theory of mind, it is believed that the function of the eye direction detector (EDD) module of individuals with ASD is impaired, resulting in their weakened ability to detect and process information. They cannot understand or speculate others' psychological state (Remington *et al*., 2009). The theory of weak central coherence holds that TD (typical development) individuals pay attention to both the global and the local information when processing the two kinds of information and can integrate them. However, the central coherence of ASD is weak, which makes it hard to extract the global information for information processing. It also affects their attention to different information, such as less attention to social information, more attention to non-social information and restricted interest objects (Müller and Frith, 2005). The theory of broken mirror is proposed based on the abnormalities of the mirror neuron system (MNS) in ASD. The MNS helps individuals to imitate others' expression in social activities. However, the MNS of individuals with ASD is damaged, and they have difficulties in imitating others' facial expressions (Wang and Chen, 2010).

Based on the above theories, researchers have started some attempts to identify children with ASD using behavioral or cognitive data. Compared with TD children, ASD children have defects in their expression imitation abilities. In particular, Rozga *et al*. (2009) identified defects in the facial mimicry responses of high-functioning ASD individuals. Likewise, Samad *et al*. (2018) analyzed expression muscles to evaluate whether ASD individuals could imitate other individuals' expressions, and they recognized spontaneous expression imitation as a behavioral marker of ASD in children. Jaiswal *et al*. (2017) developed an algorithm that used facial expression data to automatically distinguish between participants with attention deficit hyperactivity disorder (ADHD) and those with ASD. Social cognition refers to the ability to perceive and interpret social information, and ASD children also have social cognition impairments (Sasson *et al*., 2011b).

Previous studies have identified many social cognition impairments that ASD individuals frequently possess, including those that affect motion prediction (Hubert *et al*., 2007), spatial order (Sasson *et al*., 2007), target recognition (Sasson, 2006), affect recognition (Eack *et al*., 2015), and the advanced theory of mind (Baron-Cohen *et al*., 1997). Furthermore, eye fixation is commonly used to measure social preferences and evaluate social attention in ASD children. Constantino *et al*. (2017) determined that infants with ASD demonstrate atypical behavior in their preferential attention as well as in the timing, direction and targeting of their eye movements. These behaviors are strongly influenced by genetic factors.

Shaddy (2006) found that the pupil diameters of ASD children decreased when they were watching faces, while those of TD children increased. This result indicates that the ASD children, unlike the TD children, were not interested in faces. It also demonstrates that a child's pupil response can be used to distinguish between ASD and TD children. Similarly, Wang *et al*. (2018) found

that ASD children have abnormal face processing and eye-tracking skills and that analyzing these features may aid in understanding social communication disorders in ASD children, as well as provide behavioral indicators for early ASD identification. Overall, these research studies suggest that data on a child's eye fixation, facial expression, and cognitive level can be useful for distinguishing between ASD and TD children.

Notably, multi-view learning has been widely used in biomedical fields, and many achievements have been made in bioinformatics and neuroimaging regarding gene expression clustering, patient classification, brain network analysis, and biomarker identification. In bioinformatics and neuroimaging, multiple experiments can be conducted on a set of samples to obtain more types of patient data (for example, image data, physiological signals, behavioral data, or text related to the same patients), and the resulting heterogeneous data can be used to problem-solving issues with the methodology (Serra *et al*., 2018). For example, Rundo *et al.* (2017a) proposed an automated prostate gland segmentation method using multispectral T1-weighted and T2-weighted magnetic resonance imaging (MRI). In contrast to a traditional single processing pipeline applied on either T1-weighted or T2-weighted MRIs, their multi-view approach combined T1-weighted and T2-weighted MRI structural information, which significantly enhanced the prostate gland segmentation. In addition, Rundo *et al.* (2017b) developed an automated multimodal positron emission tomography (PET) imaging and MRI segmentation method designed for Gamma Knife treatments, since the joint use of MRI and PET images can convey different but complementary imaging information to enhance treatment planning.

Some studies have evaluated the use of multimodal data specifically for early ASD identification. For this purpose, Halim *et al*. (2018) combined questionnaires and home videos in their machine learning algorithm, and they obtained a significant accuracy improvement over single-modal data models. Drimalla *et al*. (2018) developed a predictive model taking the subjects' voices and facial expressions as its input. This model detected ASD in children more accurately than a single-modal approach. To investigate the behavioral markers of ASD, Samad *et al*. (2018) combined data on facial expression, visual scanning, and eye-hand coordination, and they concluded that multimodal data may provide quantitative insights into ASD that can facilitate early detection. Jaiswal *et al*. (2017) fused questionnaire, facial expression, head pose, and body posture data to identify ASD in children, and they obtained a higher identification accuracy when using multimodal rather than single-modal data.

Compared with single-modal data, multimodal data provide more features and can produce better recognition results. However, multimodal data fusion is still in its infancy in the field of ASD identification in two important ways. First, existing studies have adopted weak classifiers and traditional ensemble classifiers with low accuracies and stabilities, and the over-fitting problem can be further optimized. Second, the fusion methods in previous studies have ignored the diversity of their data sources and time synchronization, both of which affect the objectivity and accuracy of their fusion results.

For these reasons, in this study, we collected eye fixation and facial expression data from participants while they watched a short video, as well as cognitive level data from an interactive platform. After fusing the data, we then conducted early ASD identification with a multimodal framework. The main contributions of this study are as follows: (i) for the first time, data on a child's eye fixation, facial expression, and cognitive level were used for the early intelligent identification of ASD, and they were verified to be useful indicators for identifying ASD in children; (ii) different discriminative abilities of data on a child's eye fixation, facial expression, and cognitive level in ASD recognition task were explored, as well as the information complementarity of these data; (iii) an optimized random forest (RF) algorithm based on weighted decision trees was used to improve the classification accuracy, and a hybrid multimodal data fusion framework based on the data source and synchronization was developed to ensure the reliability of the classification results.

The methodology of the proposed framework is presented in Section 2. The feature extraction method is described in Section 3. The improved RF classification algorithm is presented in Section 4, wherein the details of the hybrid fusion method are also provided. The experiment results are provided in Section 5. Finally, the conclusions are given in Section 6.

2. Methodology

2.1. Data. This study was approved by our institutional review board. The data used in this study were collected from two groups of participants: (i) 50 ASD children aged 3–6 years (median = 4 years 6 months, standard deviation = 9 months) and (ii) 50 TD children aged 3–6 years (median = 4 years 8 months, standard deviation = 7 months). The ASD children were recruited from special education schools, and their diagnoses met the criteria of *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition*. Aside from ASD, these children were not diagnosed with any other disorders. The TD children were recruited from a regular kindergarten, and these children were screened to exclude any with psychiatric or neurological disorders, including ASD. There was no significant difference in age or sex between the two groups. The children's parents provided written informed

consent for their participation in this study.

The stimulus was a video clip with social and non-social information. The social information was a character with facial expressions, actions, and a voiced message. The non-social information includes backgrounds and ASD circumscribed interests, such as line-up trains and spinning wheels. We intended to analyze data reflecting the children's attention, cognitive abilities, and facial expression imitation abilities. We collected data on eye fixation and facial expression while the children watched the video and, based on the data, then analyzed the children's social attention, e.g., preferences and expression imitation abilities (Traynor *et al.*, 2019; Manfredonia *et al.*, 2018). In addition, the children's social cognitive abilities were analyzed using their responses to the social cognition questions (Kerrianne *et al.*, 2019).

2.2. Proposed framework. We developed a multimodal framework capable of automatically identifying ASD in children. In the data acquisition stage, multimodal data were collected with non-invasive sensors, including a Tobii Eye Tracker, a video camera, and a personal computer. These sensors provided information on eye fixation, facial expression, and cognitive level, respectively. Next, the features were extracted. First, the number of fixation coordinates in each cluster was extracted as an eye fixation feature using the *K*-means algorithm. Second, the number of frames containing a smiling expression in each time interval was extracted as a facial expression feature with the use of an improved facial expression recognition algorithm boosted by soft label. Finally, the answers and response times collected with an interactive question-answer platform were extracted as cognitive level features. Features with the same source and synchronization then underwent feature fusion, and an optimized RF algorithm based on weighted decision trees was applied to the classification model, which became the input for the decision fusion stage. After this stage, the final classification result was obtained. The experimental scene and the proposed framework are shown in Fig. 1.

2.3. Formal definition of identifying children with ASD. In our study, multimodal data on eye fixation, facial expression, and cognitive level were collected from the children. The identification of children with ASD was achieved using the below definition and formulas.

Definition 1. The identification of children with ASD is a learning prediction function:

$$c \to \{0, \quad 1\},$$

$$c(f_1, \ldots, f_s) = \begin{cases} 1, & \text{the child was identified ASD,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $f_1, f_2, \ldots, f_s$ represent the features extracted from the data on eye fixation, facial expression, and cognitive level.

As seen in Definition 1, the identification of ASD is binary classification based on behavioral data, physiological data, and cognitive data. Using the learning prediction function c , children with ASD can be distinguished from TD children. There are two problems to be solved: (i) how to extract features from the different data effectively and (ii) how to identify the child using those features and the fused classification results. Feature extraction is done to extract effective information from the data, and usually the features are described with a structured mathematical form. The differences between ASD and TD children can be reflected by different features. Therefore, if all the distinguishing features are fused and sent to the network for training, the accuracy of the identifications will improve. The fusion features of the j-th child can be represented as

$$I^j = \left[I_E^{(j)}, I_F^{(j)}, I_A^{(j)}, I_T^{(j)} \right], \tag{2}$$

where I represents the distinguishing feature vector, E and F represent eye fixation data and facial expression data, and A and F represent the answers to questions and response time to questions in the cognitive data, respectively. Therefore, the sample set D can be defined as:

$$D = \left\{ \left(I^1, y^1\right), \left(I^2, y^2\right), \ldots, \left(I^n, y^n\right) \right\}, \tag{3}$$

where (I^n, y^n) is the n-th training sample, and $y^n \in (0, 1)$ is the label of the n-th training sample, indicating whether or not the child has ASD.

3. Feature extraction

3.1. Eye fixation features. Numerous studies have demonstrated that ASD children have atypical attention and processing patterns for both social and non-social information (Greene *et al.*, 2011). ASD children struggle to integrate information expressed by the social cues of eye gaze, head orientation, and body orientation (Ashwin *et al.*, 2015), and are slow to detect social targets when they are in complex environments (Zhao *et al.*, 2017). These atypical attention and processing patterns are reflected in the number of fixation points in different information areas. ASD children usually have more fixation points in non-social information areas than social areas, while TD children have more fixation points in social information areas than non-social areas (Chitategmark, 2016). Moreover, autistic children give

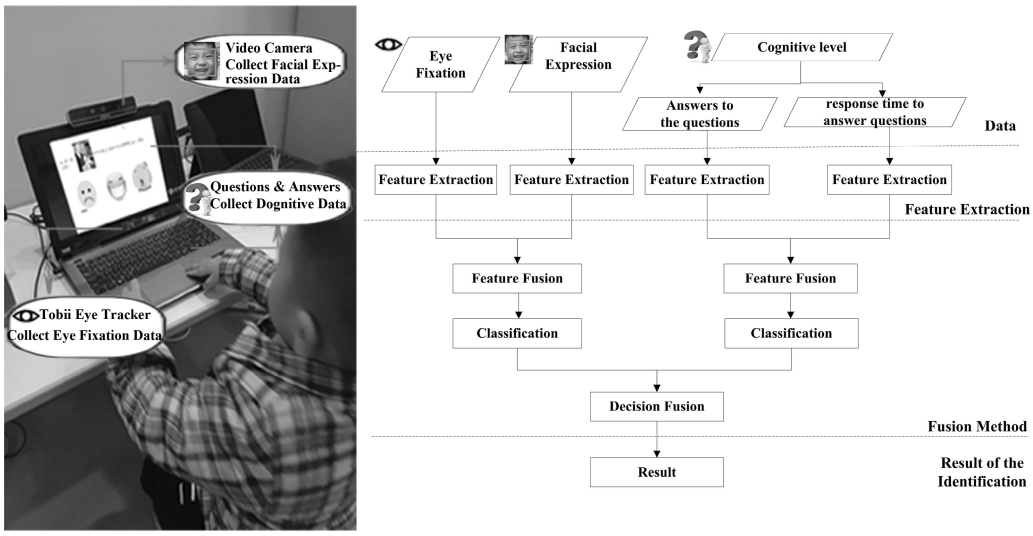


Fig. 1. Experimental scene and the proposed framework.

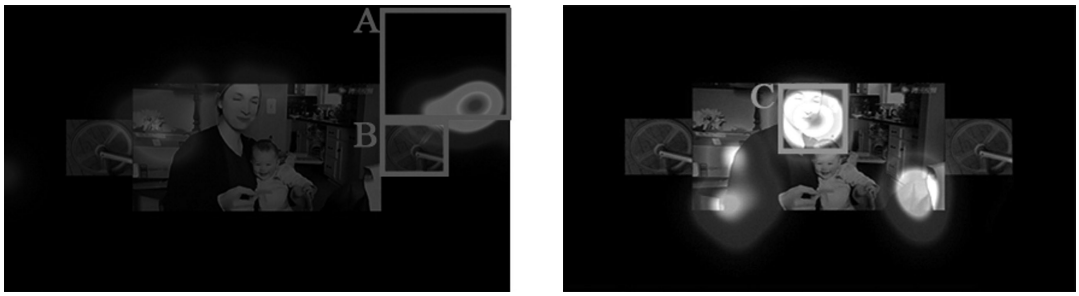
perseverative attention to objects related to their own circumscribed interest items (Sasson *et al*., 2011a).

Following these observations, we divided the video interface into different parts and counted the number of fixation points in each part. The totals were used as classification features to distinguish between ASD and TD children. Figure 2 contains a segmented video image with a social information area, a non-social information area, and a circumscribed interest area (the spinnyanging wheels). The colors of the areas represent the extent of the participants' attention.

In our study, the ASD children had a visual preference for the non-social information areas and circumscribed interest areas, while the TD children paid more attention to the social information areas. In previous studies, the interface was empirically divided into different areas of interest (AOI) (Yi *et al*., 2014). However, the AOI distribution is random and fragmentary in complex scenarios, and it is unreliable to divide these AOI without statistics. For more accurate results, we performed the k-means algorithm as outlined in a previous study (Liu *et al*., 2016), in which fixation points are clustered and divided into k clusters with distinct cluster centroids. After k cluster centroids were designated, each fixation coordinate was assigned to the cluster with the closest cluster centroid. The number of fixation points in each cluster was extracted as a feature, and k clusters corresponded to k features. Compared with the AOI, the k-means algorithm used in this study was a data-driven method, and the extracted features were consequently more accurate and objective. The features extracted from the eye fixation data can be represented as $E_1, E_2, \ldots, E_k$, such that E_1 is the number of fixation coordinates in the first cluster and E_k in the k-th cluster.

3.2. Facial expression features. Earlier studies have found significant differences in the facial expression imitation abilities of ASD and TD children (Manfredonia *et al*., 2018). The stimulus used in this study encouraged the children to spontaneously imitate the on-screen facial expressions, and the gathered data were used to evaluate their expression imitation abilities. This approach could be further studied in other fields, such as computer vision, where a facial expression recognition (FER) algorithm could be used to analyze children's facial expression imitation ability, which may be a feasible way to detect children with ASD. Owing to the content of the stimulus, we only detected the smiling expression in this study.

For our purposes, FER remains challenging due to the complexities and ambiguities in the facial expressions of ASD children, who usually exhibit a combination/mixture of emotions instead of a single emotion. Thus, traditional FER is not optimal for analyzing the facial expressions of children with ASD (Trevisan *et al*., 2018; Gan *et al*., 2019). To address this problem, our previous research on facial expression recognition based on the convolutional neural network (CNN) and the soft-label method was used to detect the facial expression of the children (Gan *et al*., 2019). A soft label can annotate multiple labels on a combination/mixture expression, thus providing a more useful description of complex expressions. The framework of the proposed facial expression detection algorithm is shown in Fig. 3. This framework mainly involved three steps: (i) a CNN model was trained using hard-label queries, (ii) soft labels were obtained by fusing the latent label probability distribution predicted by the trained model, and (iii) multiple base classifiers were trained to improve the generalization performance of the ensemble classifier. The architecture of the soft



(a) (b)

Fig. 2. Relationship between the eye fixation features and the different information areas: the eye fixation features of the ASD children (a), the eye fixation features of the TD children (b). The areas represent the different interest areas: 'A' indicates the non-social information area, 'B' the circumscribed interest area, and 'C' the social information area.

label constructor and base classifiers were initialized by a VGG16 model (Parkhi *et al*., 2015) containing 13 convolution layers and 3 fully connected layers. The convolution layers were divided into 5 groups, which were followed by 5 max pooling layers, and the resolution of the output feature maps were 1/32 of the input image. Feature maps were converted to prediction scores at the fully connected layers. However, the last fully connected layer was modified to give the final result of the expression recognition. Each section of 40 frames was used as a time interval, and the number of frames containing a smiling expression in each time interval was taken as a facial expression feature. The features can be represented as $F_1, F_2, \ldots, F_i$, such that F_1 is the number of smiling frames in the first time interval and F_i in the i-th time interval.

3.3. Cognitive level features. ASD children struggle to integrate information expressed by the social cues of eye gaze, head orientation, and body orientation (Yi *et al*., 2013), and are slow to detect social targets in complex environments (Zhong *et al*., 2019). Thus, we developed questions based on the social information in the stimulus, and the response time for each question was used to describe the different cognitive levels of the two groups. Each participant answered 6 questions after they watched the video. In total, there were 14 features: 6 answers, 6 response times, 1 total score, and 1 total response time. The interactive platform used to collect the cognitive level data is shown in Fig. 4. The feature values were normalized with min-max normalization. As shown in Table 1, the two types of cognitive features were their answers to the questions and response times for the questions.

4. Classification and the data fusion method

4.1. Classification algorithm. After feature extraction, a classification algorithm with different feature fusion layers was implemented. We developed an improved RF algorithm based on weighted decision trees. RF is an ensemble learning algorithm with a multitude of decision trees constructed by randomly sampling from the training and feature sets. RF prediction is decided by the votes from all of the decision trees to avoid over-fitting as much as possible. To reduce the instability caused by different prediction abilities among the decision trees, we evaluated the classification abilities of the decision trees according to the mutual information and assigned weights for each decision tree. Mutual information is a measure used in information theory to calculate the correlation between two variables, and it represents the uncertainty of a random variable after observing another random variable.

The process for implementing the improved RF algorithm using the weighted decision trees is described as follows:

Step 1. Create a decision tree (Kantavat *et al*., 2018). For a sample data set D, if the samples can be divided into m classes, the probability that a sample belongs to the m-th class is P_m, and the Gini index of the probability distribution is:

$$\text{Gini}(p) = 1 - \sum_{m=1}^{M} p_m^2. \tag{4}$$

A larger Gini index represents greater uncertainty; a smaller Gini index indicates higher purity.

Table 1. Data types and feature descriptions.

Data	Features	Descriptions
Answers to questions	$A_1, \ldots, A_n, A_s$	A_n is the score to the n-th question. A_s is the total score.
Response time to questions	$T_1, \ldots, T_m, T_s$	T_m is the response time to answer the m-th question. T_s is the total response time.

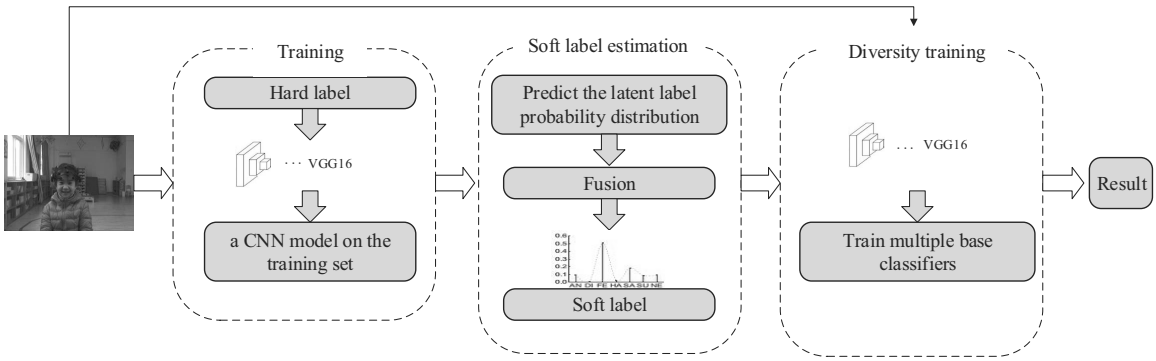


Fig. 3. Framework of the proposed facial expression detection algorithm.

The sample data set D can be divided into two classes ($D1$ and $D2$) by a feature, after which the Gini index of the data set D can be expressed as

$$\text{Gini}(D) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2) \quad (5)$$

Creating a decision tree involves creating new feature nodes with the smallest $\text{Gini}(D)$ until the decision tree stops growing.

Step 2. Create an RF. The RF is an ensemble learning algorithm with a multitude of decision trees constructed by randomly sampling from the training and feature sets. For example, $h(X, \theta_k)$ is a decision tree for $k = 1, 2, \ldots, K$, where X is a multi-dimensional vector set representing the training data and θ_k is an independent and identically distributed random vector set extracted from X. In this equation, θ_k determines the classification ability of the decision tree, and K represents the number of decision trees. The RF prediction is decided by the votes from all of the decision trees to avoid over-fitting as much as possible, and the final prediction of the RF can be given as

$$H(X) = \arg\max_{Y} \sum_{k=1}^{K} I\left(h(X, \theta_k) = y_i\right), \quad (6)$$

where $h(X, \theta_k)$ is the prediction of the k-th decision tree, Y represents the actual label vector $y_i \in Y, i = 1, \ldots, m$, and $I(\cdot)$ represents the indicative function.

Fig. 4. Interactive platform that collected cognitive level data.

Step 3. Assign a weight to each decision tree based on the mutual information.

Z_j represents the vector of the j-th feature of the training samples. The mutual information of Z_j and Y can be used to calculate the influence of a feature on the final result through the following equation:

$$\begin{aligned} I(Z_j; Y) &= \iint P(Z_j, Y) \log \frac{P(Z_j, Y)}{P(Z_j)P(Y)} \, \mathrm{d}z \, \mathrm{d}y \\ &= H(Y) - H(Y|Z_j), \end{aligned} \quad (7)$$

where $P(Z_j, Y)$ represents the joint probability distribution of Z_j and Y, $P(Z_j)$ represents the marginal distributions of Z_j, $P(Y)$ represents the marginal distributions of Y, $H(Y)$ represents the entropy of Y, and $H(Y|Z_j)$ represents the entropy of Y after the given variable Z_j.

The voting weight of a decision tree should be the sum of the mutual information of all of its features and can be defined as follows:

$$p = \alpha \sum_{j=1}^{J} I(Z_j; Y), \quad (8)$$

where $I(Z_j; Y)$ represents the mutual information of the Z_j and Y, J represents the number of features, and α represents the normalization factor of the voting weight.

Step 4. The final prediction of the RF based on the weighted decision trees.

$H = \{h(X, \theta_1), h(X, \theta_2), \ldots, h(X, \theta_k)\}$ represents the RF model, and the final prediction of the model is

determined as follows:

$$\max\left\{c|c_i=\sum_{k=1}^{K}p_kI(h(X,\theta_k)=y_i),\ y_i\in Y,i=1,\ldots,m\right\} \quad (9)$$

where p_k is the voting weight of the k-th decision tree and $I(\cdot)$ represents the indicative function. If the prediction of $h(X,\theta_k)$ is y_i, the value of $I(\cdot)$ is 1. Otherwise, it is 0. Here c_i is the weighted voting result, and the final prediction of the model is the maximum c_i. This improved RF algorithm based on weighted decision trees is shown in Fig. 5.

4.2. Hybrid fusion method based on the data source and time synchronization. Multimodal data fusion processes multiple information sources with different features and time series. There are three fusion methods: feature fusion, decision fusion, and hybrid fusion (Poria *et al*., 2017). For feature fusion, the features extracted from various modalities are fused as a general feature vector, which was analyzed for a final result. In decision fusion, the features of each modality are classified independently, and the classification results of different modalities were identified as sub-decisions, which were fused as a decision vector to gain the final result. The advantage of feature fusion is that the complementary information between various multimodal features was used at an early stage, and can potentially provide a better task accomplishment, but the final classification result is not reliable if a modality is lost or wrong. Decision fusion is a more robust approach that combines the sub-decisions of each modality, but it does not take advantage of complementary information at the early stage. Hybrid fusion combines the advantages of feature fusion and decision fusion, and it can flexibly and easily fuse multi-source asynchronous data. In this study, the data were collected from different sensors, and the data collection time was asynchronous. Hence, the hybrid fusion process was divided into two levels. The data with the same source and synchronization were fused in the first level for feature fusion, and then the results were fused in the second level for decision fusion. The multimodal data fusion method is shown in Fig. 6.

In Fig. 6, we fused the data from the same source and collection time. The behavioral data on eye fixation and facial expression were collected synchronically, as were the cognitive data on answers and response times.

In the first fusion level, the eye fixation feature vector and the facial expression feature vector were connected and were sent into the feature pool of RF1 for feature selection. Similarly, the feature vectors of answers and response times were connected and sent into the feature pool of RF2 for feature selection. As a result, the behavioral data were fused by RF1 and the cognitive data were fused by RF2. To improve the prediction abilities of our study, we used the improved RF algorithm (RF1 and RF2) to assign weights to each decision tree. Those weights were calculated according to the mutual information, as outlined in Eqn. (8). The classification method was given in Eqn. (9).

In the second fusion level, we fused the decisions of RF1 and RF2. The decision fusion weight of RF1 was the sum of the mutual information of all of its decision trees, and the decision fusion weight of RF2 was calculated in the same way. Then, the decisions and their fusion weights were multiplied respectively, and the results were added and assigned to R. Compared with the threshold, if R was greater than or equal to the threshold, the final result was 1; otherwise, the final result was 0. The threshold was obtained by traversing all R. For each traversal, the threshold was set to the current R, and the classification accuracy was calculated under this threshold setting. After all traversals, all of the classification accuracies were compared, and the threshold corresponding to the highest classification accuracy was set as the final threshold.

5. Results

5.1. Validation method. During the experiment, each participant (50 ASD children and 50 TD children) generated a sample containing data on eye fixation, facial expression, and cognitive level. In a machine learning framework, the data set is usually divided into the test set and the training set. The training set is used to train the model, while the test set is used to evaluate its ability to generalize (Al-Jarrah *et al*., 2015). To provide the model with enough training samples, we used the leave-one-out cross-validation method (Xu *et al*., 2018). If the size of the data set D was n, $n-1$ samples were used for training and the remaining one sample was used for testing. One sample was taken from D and added to the test set until all of the samples were tested, and then the test accuracy average was calculated as the final result.

5.2. Analysis of time complexity. The improved RF algorithm proposed in this paper used decision trees as the base classifier, where the number of base classifiers was K and the number of samples was n. Two random forest models were used to fuse the behavioral features and cognitive features, respectively. The feature dimension selected by the RF algorithm was z. In the experiment, the leave-one-out cross-validation method was performed. Thus, each kind of fusion feature needed to run the RF algorithm n times. In the process of constructing the decision tree, the growth of the tree was not pruned, so the time of training each base classifier was less than $O(nz\log n)$. Furthermore, the weights of the

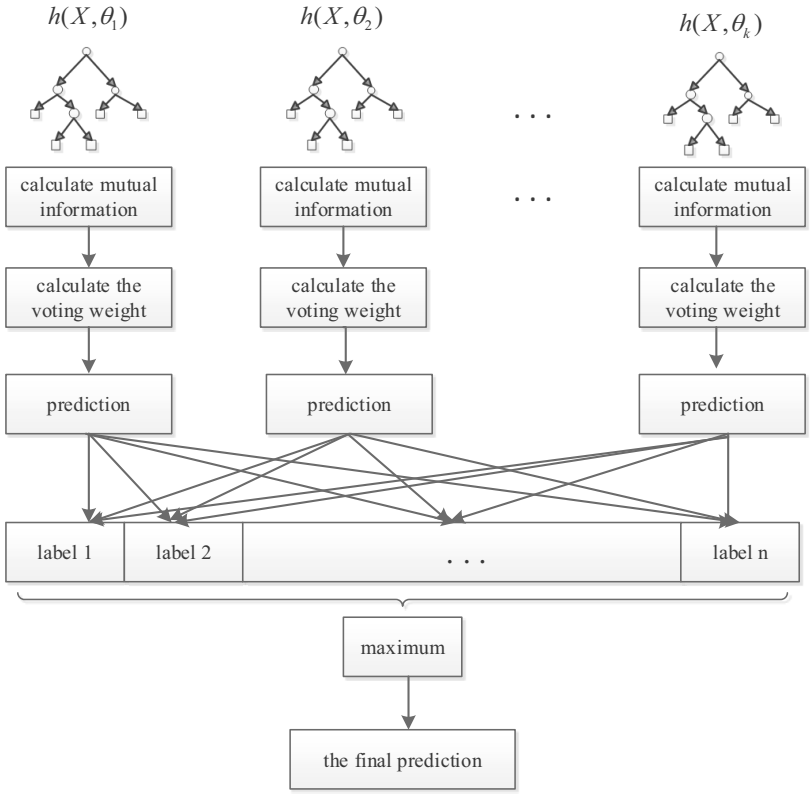


Fig. 5. Improved RF algorithm.

decision tree were calculated by the mutual information of samples' features and labels, and the time complexity of the mutual information was $O(z)$. Therefore, the time complexity of the improved RF algorithm for each fusion feature was $O(Kz(1+n\log n))$, and the total time complexity of the algorithm was $O(2Kz(n+n^2\log n))$.

5.3. Performance of the proposed framework. As shown in Table 2, we compared the classification accuracies of different classifiers for decision fusion and hybrid fusion, respectively. The best classification accuracy was 91%, and it was obtained with the improved RF algorithm and the hybrid fusion method proposed in this study. The best accuracies were 91% for hybrid fusion and 87% for decision fusion. The average accuracies were 86.25% for hybrid fusion and 82% for decision fusion. Notably, the hybrid fusion method proposed in this paper had the highest individual and average values. Therefore, the results indicate that the proposed framework can effectively separate ASD and TD children.

Meanwhile, we compared the classification accuracies of single-modal and multimodal fusion classification methods, as shown in Table 3. For single-modal classification (eye fixation, facial expression, answers, or response time), the maximum classification accuracy was 75% with the data of response time. The proposed hybrid fusion method had a stronger performance (accuracy 91%) than the decision fusion method (accuracy 87%).

5.4. Complementary characteristics of different data modalities. For ASD identification, we obtained a classification accuracy of 69% using only the data of eye fixation, and it was 66% using the data of facial expression, and 74% using the data of answers, and 75% using response time. For hybrid fusion based on the data source and time synchronization, there were two stages: (i) feature fusion in the first fusion level and (ii) decision fusion in the second fusion level.

Table 2. Accuracies of the different classifiers and fusion methods (%).

Classifier	Decision fusion	Hybrid fusion
RF	82	86
SVM	78	83
DA	81	85
improved RF algorithm	87	**91**
AVG	82	86.25

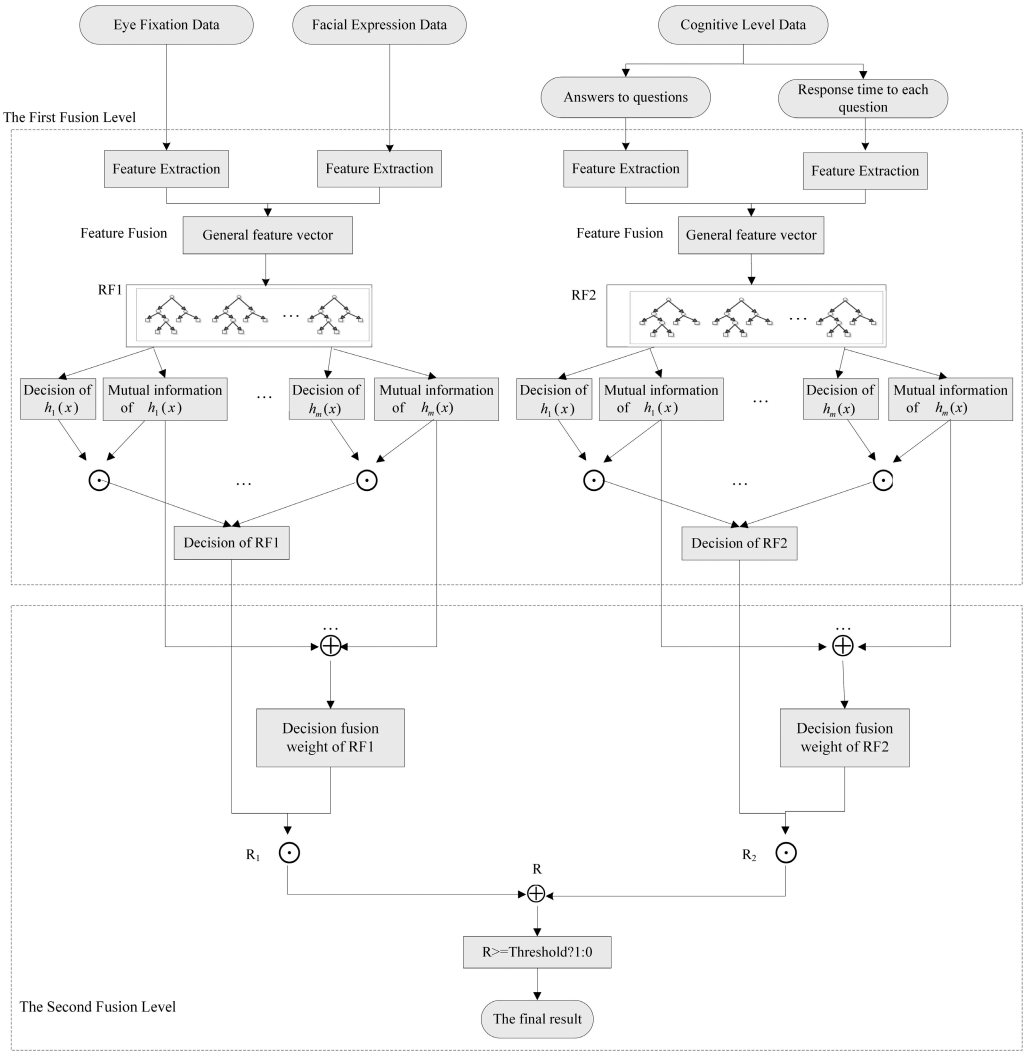


Fig. 6. Multimodal data fusion method. RF1 and RF2 are two random forests used for feature selection: here $h_m(x)$ represents a decision tree in the random forest, $\bigoplus$ is the arithmetic operator used to perform the addition operation, and $\odot$ is the multiplication operator.

Using hybrid fusion, we obtained an accuracy of 91%, which was significantly greater than that using single modality, indicating that hybrid fusion may combine the complementary information of single modality and effectively enhance the classification performance.

Table 3. Accuracies of single-modal and fusion classification methods(%).

Accuracy(%)	Eye fixation	Facial expression	Answers	Response time (%)
Single-modal	69	66	74	75
Decision fusion	87			
Hybrid fusion	**91**			

To further investigate the complementary characteristics of different data modalities, we analyzed the confusion matrices of eye fixation classification, facial expression classification, answer classification and response time classification, which could reveal the advantages and weakness of each modality. The confusion matrices of each data modality were shown in Fig. 7. We observed that eye fixation has the advantage of classifying ASD (82%) compared with facial expression (58%), whereas facial expression outperforms eye fixation in recognizing TD (74% versus 56%). It is difficult to recognize TD using only eye fixation and ASD using only facial expression, and the advantages of eye fixation and facial expression are complementary information to each other to improve the identification accuracy.

Meanwhile, it can also be seen that response time has the advantage of classifying ASD (88%) compared with answer (72%), whereas answer outperforms response time in recognizing TD (76% versus 62%), indicating that there is complementary information between the data of answers and response time.

Moreover, the misclassifications of each data modality are different. Eye fixation misclassifies more TD as ASD (44%), whereas facial expression misclassifies more ASD as TD (42%). Answers misclassify more ASD as TD (28%), while response time misclassifies more TD as ASD (38%). These results indicate that eye fixation, facial expression, answers and response time have different discriminative powers for recognition ASD and TD, and they have important complementary characteristics. As shown in Table 3, combining the complementary information of them, hybrid fusion can significantly improve the classification accuracies (91%).

6. Conclusions

Early identification of ASD in children can dramatically improve their prognosis and greatly benefit their treatment. We have presented an intelligent multimodal framework to identify ASD in children. Our main contributions are threefold. First, we have used a novel combination of eye fixation, facial expression, and cognitive level data for early ASD identification, and they were verified to be useful indicators for identifying ASD in children. Second, different discriminative abilities of data on a child's eye fixation, facial expression, and cognitive level in ASD recognition task were explored, as well as the information complementarity of these data. Third, we have presented an optimized random forest algorithm and a multimodal data fusion framework that uses a hybrid fusion method based on the data source and synchronization to ensure the reliability of the classification results. Our results indicate that data on a child's eye fixation, facial expression, and cognitive level can be useful for identifying ASD in children. The proposed framework can separate ASD children from TD children and, consequently, can facilitate early ASD identification and support intervention programs for ASD children.

This study has some limitations. First, the number of samples used was relatively small. Increasing this number would improve the accuracy and stability of the algorithm, although it is difficult to increase the number of ASD children sampled. If the number of TD children sampled is too high, the positive and negative samples will be imbalanced, thus affecting the results. In a future study, we can apply a penalty weight method to solve this problem (Yi *et al.*, 2013; Zhong *et al.*, 2019). Second, we only combined data on eye fixation, facial expression, and cognitive level, ignoring other data modalities such as EEG, peripheral physiological signals, and body movements (Halim *et al.*, 2018). In a future study, we plan to compare the identification abilities of different data modalities to construct a more comprehensive and effective framework.

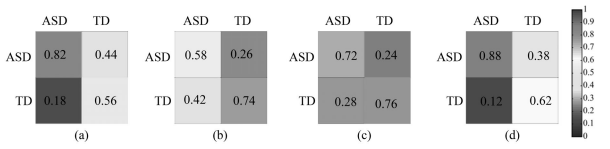


Fig. 7. Confusion matrices of single modality classification and hybrid fusion classification. Each row of the confusion matrix represents a predicted class and a column represents the target class. The element (i, j) is the percentage of samples in class j that is predicted as class i: eye fixation (a), facial expression (b), answers (c), response time (d).

Acknowledgment

This work was supported by the National Key Research and Development Program of China under the grant 2018YFB1004504, the National Natural Science Foundation under the grants 61977027 and 61807014, the Hubei Province Technological Innovation Major Project under the grant 2019AAA044, Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE under the grants CCNU19Z02002 and CCNU18KFY02, the Humanities and Social Sciences Program of the Education Department of the Henan Province under the grant 2020-ZDJH-295.

References

Achenbach, T. and Rescorla, L. (2000). *Manual for the ASEBA Preschool Forms & Profiles*, University of Vermount, Burlington, VA.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K. (2015). Efficient machine learning for big data: A review, *Big Data Research* **2**(3): 87–93.

Amaral, D.G., Schumann, C.M. and Nordahl, C.W. (2008). Neuroanatomy of autism, *Trends in Neurosciences* **31**(3): 137–145.

Ashwin, C., Hietanen, J.K. and Baron-Cohen, S. (2015). Atypical integration of social cues for orienting to gaze direction in adults with autism, *Molecular Autism* **6**(1): 5–14.

Baron-Cohen, S., Jolliffe, T., Mortimore, C. and Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome, *Journal of Child Psychology and Psychiatry* **38**(7): 813–822.

Bernier, R., Mao, A. and Yen, J. (2011). Diagnosing autism spectrum disorders in primary care, *Practitioner* **255**(1745): 27–30.

Chitategmark, M. (2016). Social attention allocation in ASD: A review and meta-analysis of eye-tracking studies, *Review Journal of Autism & Developmental Disorders* **3**(3): 209–223.

Christensen, D.L., Baio, J., Braun, K.V.N., Bilder, D., Charles, J., Constantino, J.N., Daniels, J., Durkin, M.S., Fitzgerald, R.T. and Kurziusspencer, M. (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2012, *Morbidity and Mortality Weekly Report—Surveillance Summaries* **65**(3): 1–23.

Constantino, J.N., Kennon-Mcgill, S., Weichselbaum, C., Marrus, N. and Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism, *Nature* **547**(7663): 340–344.

Drimalla, H., Landwehr, N., Baskow, I., Behnia, B. and Scheffer, T. (2018). Detecting autism by analyzing a simulated social interaction, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Berlin, Germany*, pp. 193–208.

Durkin, M., Maenner, M.J., Meaney, F.J., Levy, S.E. and DiGuiseppi, C. (2010). Socioeconomic inequality in the prevalence of autism spectrum disorder: Evidence from a US cross-sectional study, *PLoS ONE* **5**(7): e11551.

Eack, S.M., Mazefsky, C.A. and Minshew, N.J. (2015). Misinterpretation of facial expressions of emotion in verbal adults with autism spectrum disorder, *Autism* **19**(3): 308–315.

Gan, Y.L., Chen, J.Y. and Xu, L.H. (2019). Facial expression recognition boosted by soft label with a diverse ensemble, *Pattern Recognition Letters* **125**(4): 105–112.

Greene, D.J., Colich, N., Iacoboni, M., Zaidel, E., Bookheimer, S.Y. and Dapretto, M. (2011). Atypical neural networks for social orienting in autism spectrum disorders, *Neuroimage* **56**(1): 354–362.

Halim, A., Ford, G., Eric, G. and Wall Dennis, P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video identification, *Journal of the American Medical Informatics Association* **25**(8): 1000–1007.

Hubert, B., Wicker, B., Moore, D.G., Monfardini, E., Duverger, H., Da Fonseca, D. and Deruelle, C. (2007). Brief report: Recognition of emotional and non-emotional biological motion in individuals with autistic spectrum disorders, *Journal of Autism and Developmental Disorders* **37**(7): 1386–1392.

Jaiswal, S., Valstar, M.F., Gillott, A. and Daley, D. (2017). Automatic detection of ADHD and ASD from expressive behaviour in RGBD data, *IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, USA*, pp. 762–769.

Jiang, M., Sunday, M., Francis and Srishyla, D. (2019). Classifying individuals with ASD through facial emotion recognition and eye-tracking, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany*, pp. 6063–6068.

Kantavat, P., Kijsirikul, B., Songsiri, P., Fukui, K.-I. and Numao, M. (2018). Efficient decision trees for multi-class support vector machines using entropy and generalization error estimation, *International Journal of Applied Mathematics & Computer Science* **28**(4): 705–717, DOI: 10.2478/amcs-2018-0054.

Kerrianne, E., Morrison, A.E. and Pinkham, S.K. (2019). Psychometric evaluation of social cognitive measures for adults with autism, *Autism Research* **12**(5): 766–778.

Liu, W., Li, M. and Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework, *Autism Research* **9**(8): 888–898.

Manfredonia, J., Bangerter, A., Manyakov, N.V., Ness, S., Lewin, D., Skalkin, A., Boice, M., Goodwin, M. S., Dawson, G. and Hendren, R. (2018). Automatic recognition of posed facial expression of emotion in individuals with autism spectrum disorder, *Journal of Autism and Developmental Disorders* **27**(10): 1–15.

Müller and Frith, U. (2005). Autism-explaining the enigma, *Kindheit Und Entwicklung* **14**(4): 257.

Parkhi, O., Vedaldi, A. and Zisserman, A. (2015). Deep face recognition, *British Machine Vision Conference, Swansea, UK,* p. 6.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion, *Information Fusion* **37**(2): 98–125.

Remington, A., Swettenham, J., Campbell, R. and Coleman, M. (2009). Selective attention and perceptual load in autism spectrum disorder, *Psychological Science* **20**(11): 1388–1393.

Rozga, A., Mumaw, M., King, T. and Robins, D.L. (2009). Lack of emotion-specific facial mimicry responses among high-functioning individuals with an autism spectrum disorder (poster), *International Meeting for Autism Research, Chicago, IL, USA*, pp. S43–S44.

Rundo, L., Militello, C., Russo, G., Garufi, A., Vitabile, S. and Gilardi, M. (2017a). Automated prostate gland segmentation based on an unsupervised fuzzy c-means clustering technique using multispectral T1w and T2w MR imaging, *Information* **8**(2): 1–28.

Rundo, L., Stefano, A., Militello, C., Russo, G., Sabini, M.G. and Arrigo, C. (2017b). A fully automatic approach for multimodal PET and MR image segmentation in Gamma Knife treatment planning, *Computer Methods & Programs in Biomedicine* **144**(3): 77–96.

Samad, M.D., Diawara, N., Bobzien, J.L., Harrington, J.W., Witherow, M.A. and Iftekharuddin, K.M. (2018). A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response

data, *IEEE Transactions on Neural Systems & Rehabilitation Engineering* **PP**(99): 1–1.

Sasson, N.J. (2006). The development of face processing in autism, *Journal of Autism & Developmental Disorders* **36**(3): 381–394.

Sasson, N.J., Elison, J.T., Turner-Brown, L.M., Dichter, G.S. and Bodfish, J.W. (2011a). Brief report: Circumscribed attention in young children with autism, *Journal of Autism & Developmental Disorders* **41**(2): 242–247.

Sasson, N.J., Pinkham, A.E., Carpenter, K.L. and Belger, A. (2011b). The benefit of directly comparing autism and schizophrenia for revealing mechanisms of social cognitive impairment, *Journal of Neurodevelopmental Disorders* **3**(2): 87–100.

Sasson, N., Tsuchiya, N., Hurley, R., Couture, S.M., Penn, D.L., Adolphs, R. and Piven, J. (2007). Orienting to social stimuli differentiates social cognitive impairment in autism and schizophrenia, *Neuropsychologia* **45**(11): 2580–2588.

Serra, A., Galdi, P. and Tagliaferri, R. (2018). Machine learning for bioinformatics and neuroimaging, *Wiley Interdisciplinary Reviews: Data Mining & Knowledge Discovery* **8**(5): e1248.

Shaddy, D.J. (2006). Visual scanning and pupillary responses in young children with autism spectrum disorder, *Journal of Clinical & Experimental Neuropsychology* **28**(7): 1238–1256.

Tariq, Q., Daniels, J. and Schwartz, J.N. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study, *PLoS Medicine* **15**(11): e1002705.

Traynor, J.M., Gough, A., Duku, E., Shore, D.I. and Hall, G.B.C. (2019). Eye tracking effort expenditure and autonomic arousal to social and circumscribed interest stimuli in autism spectrum disorder, *Journal of Autism and Developmental Disorders* **49**(1): 1988–2002.

Trevisan, D.A., Hoskyn, M. and Birmingham, E. (2018). Facial expression production in autism: A meta-analysis, *Autism Research* **11**(2): 1586–1601.

Wang, G.S., Chen, J.Y. and Zhang, K. (2018). The perception of emotional facial expressions by children with autism using hybrid multiple factorial design and eye-tracking, *Chinese Science Bulletin* **63**(31): 3204–3216, (in Chinese).

Wang, Y. and Chen, W. (2010). Broken mirror theory of autism, *Advances in Psychological Science* **18**(2): 297–305.

Xu, L., Fu, H.Y., Goodarzi, M., Cai, C.B., Yin, Q.B., Wu, Y., Tang, B.C. and She, Y. B. (2018). Stochastic cross validation, *Chemometrics & Intelligent Laboratory Systems* **175**(4): 74–81.

Xu, M., Shen, J. and Yu, H.Y. (2017). A review on data-driven healthcare decision-making support, *Industrial Engineering and Management* **21**(1): 1–13.

Yi, H., Song, X.F., Jiang, B., Liu, Y.F. and Zhou, Z.H. (2013). Fault diagnosis based on self-tuning support vector machine in sample unbalance condition, *Transactions of Beijing Institute of Technology* **33**(4): 394–398.

Yi, L., Feng, C., Quinn, P.C., Ding, H., Li, J., Liu, Y. and Lee, K. (2014). Do individuals with and without autism spectrum disorder scan faces differently? A new multi-method look at an existing controversy, *Autism Research* **7**(1): 72–83.

Zhao, S., Uono, S., Yoshimura, S., Kubota, Y. and Toichi, M. (2017). Atypical gaze cueing pattern in a complex environment in individuals with ASD, *Journal of Autism Developmental Disorders* **47**(7): 1978–1986.

Zhong, S.S., Li, X. and Zhang, Y.J. (2019). Fault diagnosis of civil aero-engine driven by unbalanced samples based on DBN, *Journal of Aerospace Power* **34**(3): 708–716.

Zunino, A., Morerio, P. and Cavallo, A. (2018). Video gesture analysis for autism spectrum disorder detection, *24th International Conference on Pattern Recognition (ICPR), Beijing, China*, pp. 3421–3426.

Zwaigenbaum, L., Bryson, S., Lord, C., Rogers, S., Carter, A., Carver, L., Chawarska, K., Constantino, J., Dawson, G. and Dobkins, K. (2009). Clinical assessment and management of toddlers with suspected autism spectrum disorder: Insights from studies of high-risk infants, *Pediatrics* **123**(5): 1383–1391.

Jingying Chen received her BS and MS degrees from the Huazhong University of Science and Technology, Wuhan, China, and her PhD degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001. She has worked as a post-doc in INRIA, France, and a research fellow with the University of St Andrews and the University of Edinburgh, UK. She is currently a professor with the National Engineering Center for E-Learning, Central China Normal University, China. Her research interests include information technology in education, computer vision, pattern recognition and computer-human interaction.

Mengyi Liao received her BS degree in the School of Computer and Information Engineering at Henan University, Kaifeng, China, in 2006, and her MS degree in the School of Computer and Information Engineering at Xidian University, Xi'an, China, in 2009. She is currently pursuing her PhD degree at the National Engineering Research Center for E-Learning, Central China Normal University. Her research interests include information technology in education, special education and machine learning.

Guangshuai Wang is currently a doctoral candidate at the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include information technology for special education, and serious games for the assessment for autism spectrum disorder.

Chang Chen is currently a doctoral candidate at the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include information technology for special education, and machine learning.

Received: 25 October 2019
Revised: 27 March 2020
Re-revised: 3 June 2020
Accepted: 4 June 2020