A deep learning algorithm-based visual strategy intervention study for children with autism spectrum disorders – extraction and detection of children's behavioral features

Y. LIN

Department of Applied Psychology, Lingnan University, Tuen Mun, New Territories, Hong Kong, China

Abstract. – OBJECTIVE: Autism spectrum disorder is a group of neurodevelopmental disorders. The disease's etiology is unclear, and there is no specific drug treatment for the core symptoms of autism spectrum disease. The study aims to explore effective intervention methods for children with autism spectrum disorders.

MATERIALS AND METHODS: This paper proposes a visual strategy intervention method for children with autism spectrum disorders. This method combines feature extraction and abnormal behavior detection and can use a visual cue strategy to integrate children into social groups. Firstly, the spatial-temporal feature fusion structure is added to extract behavioral features from children, and the spatial information contained in MotionNet is fused with temporal features. Optical Flow Feature (OFF) subnetwork is added to the optical flow extraction feature network. Each layer feature is input to the OFF subnet to extract the time feature further. Then, a behavior detection method based on the sequential pool is proposed. This method combines attention mechanism and clustering pool to effectively describe human behavior dynamics in the long, redundant video under complex background. Finally, feature extraction and behavior detection experiments are carried out on SDUFall, Weizmann, and HMDB51 data sets.

RESULTS: The model's accuracy is still slightly higher than others in that only the video Red-Green-Blue (RGB) frame is used as input. Compared with OFF, SDUFall can reach 88.64%, and HMDB51 can only reach 63.81%. In contrast, the proposed model can reach 72.09%, higher than others. The descriptor obtained the best result of 92.57%, which is 3.64%, 2.58%, and 1.73% higher than the other three comparison descriptors. The data show that the method presented here is effective and has advantages in detecting children's abnormal behavior.

CONCLUSIONS: This method and visual intervention for children with autism spectrum disorders can help them to overcome social barriers.

Key Words:

Deep learning, Autism, Feature extraction, Abnormal behavior detection, Visual strategy intervention.

Introduction

Autism Spectrum Disorder (ASD), also known as autism, is one of the most common and serious developmental disorders in children with social interaction disorders, narrow interests and stereotyped behaviors, and varying degrees of psychiatric problems, such as attention deficit, hyperactivity, sleep disorders, disruptive and impulsive behavior control disorders, and obsessive-compulsive disorder, which usually appear in early childhood. These deficits usually appear in early childhood development and can affect patients throughout their lives. In recent years, the prevalence of autism has been increasing significantly, from 13.4/1,000 in 2010 to 15.3/1,000 in 2012 and 17.0/1,000 in 2014, and has become a severe public health problem worldwide. The second national sample survey of people with disabilities in China found that people with ASD accounted for 22.34% of those with mental disabilities aged 2 to 6 years. A recent survey[1] showed that the prevalence of autism in China was 1%. According to the survey, the risk of death of autistic patients is 2.8 times higher than that of patients with other diseases of the same sex and age. 45% of autistic patients have intellectual disabilities, 32% have degenerative symptoms, and more than 70% of autistic patients have psychiatric and psychological problems of different degrees. They have a poor prognosis regarding education, employment, and self-care and require long-term parental care.

The key to the treatment of autism is early detection, diagnosis, and intervention. There is

Corresponding Author: Yibing Lin, MD; e-mail: lyibing3366@126.com

growing evidence that interventions starting at the age of 2-4 years are more effective than interventions after the age of 4 years and that appropriate interventions at an early age can reduce or even prevent the development of serious behavioral problems in children later in life[2]. To date, the etiology of ASD remains an unanswered question worldwide. There is a consensus that patients with ASD exhibit a variety of developmental disorders, primarily due to biological factors in the brain. These factors include genetic factors, maternal and perinatal biological factors, infectious factors, immune factors, structural or functional brain abnormalities, neuroendocrine and neurotransmitter factors, other physiological factors, and psychosocial factors[3]. Although the specific etiology is unclear, literature on the subject proposes ASD as a neurodevelopmental disorder through the extensive neuropathological and physiological aspects performed, combined with imaging.

Currently, there is no specific systematic clinical treatment for ASD. The international treatment for ASD includes behavior modification, special education training and medication, and comprehensive treatment. However, medication cannot shorten the whole course of the disease, and its primary purpose is to control and improve the child's clinical symptoms. Meanwhile, behavior modification and education training mainly target the social-emotional disorder of the child. However, effective rehabilitation training can significantly improve the symptoms and self-care ability of children with ASD[4]. The mainstream rehabilitation training for children with ASD is shown in Figure 1. It includes: 1) behavior analysis therapy; 2) music therapy; 3) structured education therapy; 4) conductive education therapy; 5) interpersonal and developmental intervention therapy; 6) picture exchange communication therapy; 7) pharmacy therapy; 8) family intervention therapy.

In addition to the problems of effectiveness and applicability of theoretical and technical methods, future research should focus more on the combination and joint application of various methods in the intervention of children with autism. The integration of early educational interventions in United States, Hong Kong, Macao, and Taiwan medical institutions and general education in the family is more in line with the current actual situation of children with ASD in China[5]. This model is a shift from an early emphasis on the importance of child-centered institutional education to a family-centered concept. Applying family intervention therapy with other therapies has more prospects for future applications in rehabilitating children with ASD.

Based on the existing intervention therapies for children with ASD, some researchers[6-9] have enhanced them by combining artificial intelligence techniques. Deep learning methods have made various breakthroughs in behavioral feature ex-

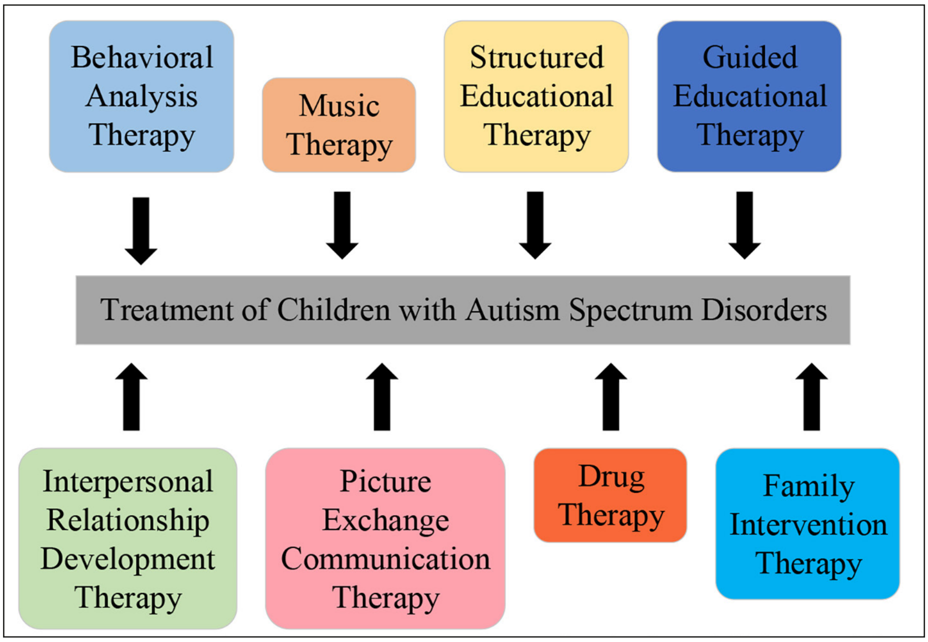
Behavioral Analysis Therapy
Music Therapy
Structured Educational Therapy
Guided Educational Therapy
Treatment of Children with Autism Spectrum Disorders
Interpersonal Relationship Development Therapy
Picture Exchange Communication Therapy
Drug Therapy
Family Intervention Therapy

Figure 1. Mainstream treatment methods for children with ASD.

traction and detection for children with ASD, and their speed and accuracy rates are much faster than traditional methods. Although IDT (Integrated Digital Terminal) and the boosting algorithms have achieved quite good results on the video behavior recognition task, they require huge computational effort[10]. Therefore, manual features are gradually being replaced by neural network features. Deep learning-based algorithms mainly use neural networks such as two streams, LSTM (Long Short Term Memory), and C3D (Convolutional three-dimensional), to model the input continuous Red-Green-Blue (RGB) frames, analyze the behavioral features in the video, and classify them based on the behavioral features. However, to pursue higher classification accuracy, the classification speed of the model is often significantly reduced. How to improve the feature extraction and behavior classification models based on deep learning methods, enhance the feature extraction speed, and apply them to visual strategy interventions for children with ASD is the key research direction.

For the existing intervention treatment methods for children with ASD, combined with deep learning, a method is reported to assist visual intervention strategies for children with ASD through feature extraction and behavioral detection. First, a new feature extraction network structure is proposed, which uses a neural network for computing optical flow instead of directly using optical flow as input. Spatial features are generated by the neural network's optical flow computation. Rich spatial-temporal features can be obtained by fusing this part of spatial features with motion features in the latter part of the temporal flow. Then, to address the problem of redundant frames affecting behavior classification accuracy, a trajectory attention map without network training is proposed. Based on this an order pool is added to construct a descriptor, which can effectively locate the motion target region in the video, and it is robust to the surrounding environment changes. It is experimentally verified that the proposed method can detect abnormal behaviors in children with ASD and provide a reference for visual strategy intervention treatment methods.

Materials and Methods

Current Status of the Application of Artificial Intelligence Technology in the Treatment of Children with ASD

Artificial intelligence technology is a newly applied technology to the developmental-behavioral rehabilitation and care of children with autism spectrum disorders. It has good applicability in improving children's developmental behavior and enhancing their quality of survival while improving the level of care and reducing the burden on caregivers[11]. This review provides a comprehensive analysis of relevant domestic and international literature. It summarizes the progress of the application of artificial intelligence technology in rehabilitating cognitive and social-behavioral functions in children with autism spectrum disorders. It provides a reference for better-targeted rehabilitation and care and has implications for the subsequent introduction of artificial intelligence technology into clinical rehabilitation.

The number of children with autism spectrum disorders continues to increase. The long intervention period and the shortage of nursing specialists and rehabilitation teachers have become increasingly prominent and have attracted widespread attention. Children with autism spectrum disorders are often afraid to communicate with others or cannot understand normal communication timely[12]. Artificial intelligence technology machines with predictability and repeatability make children with autism spectrum disorders feel more secure than human caregivers and rehabilitators.

Children can gain trust more quickly than caregivers and rehabilitators using Artificial Intelligence (AI) technology products in interventions. Besides, children can follow its guidance more efficiently than adults. AI technology machines are more likely to substantially affect the child, as they are more attractive to the child and support multi-modal interactions, including gesture, speech, and touch. Further, AI technology machines can be used not only as companions but also as toys[13]. They can be programmed to adapt to the severity of the child's autism spectrum disorder and unique behaviors according to their specific needs, thus allowing for individualized care and rehabilitation. In addition, the stability of AI technology, which is different from that of health care professionals and rehabilitators, the durability of long-term applications, and the replicability of intelligent machines make its application less costly and with the same or better rehabilitation results[14]. Additionally, the application of AI technology allows for the scientific systematization and precision of rehabilitation content. It facilitates the collection of indicators and data for assessment, clinical feedback, or research, which is very important for developing research in autism spectrum disorders.

AI-based robots and related devices provide a relatively simplified environment for children with autism spectrum disorders. They can gradually increase the complexity of cognitive changes and social behaviors children must face. Although domestic AI technologies are developing rapidly and approaching world leadership, and research on the application of robotic interventions for children with autism spectrum disorders is emerging one after another, research related to the original design of medically intelligent machines for use with children with autism is still relatively recent[15]. There is a wide variety of AI technology applications abroad, often applying different AI machines for different rehabilitation purposes.

However, there are still some problems in applying AI technology in rehabilitation. Most studies[16,17] have not systematically evaluated the generalizability of AI technology applied to the rehabilitation of autism spectrum disorders, and most studies[18,19] have small sample sizes, which are expected to be verified by more studies in the future. It is worth noting that the ultimate goal of rehabilitation is the increased ability of children with autism spectrum disorders to interact in social settings, not just limited to communication with intelligent machines.

Current Status of Research on Feature Extraction Based on Deep Learning

Generally speaking, the design goals of feature extraction usually include ignoring redundant information in the image and recording the required necessary information. Usually, such a description can be a mathematical representation such as a scalar, vector, or matrix. In this process, the key point is the extreme value point located, and the computed description vector is the feature.

The pixel points with unique characteristics such as large gradient, curvature, and brightness variation are called key points, also known as "feature points" and "interest points"; the descriptive expressions extracted from the area near the key points are called key points The features of the key points are also called "descriptors". The key points are designed to be robust to changes in brightness, rotation, and observation points and are designed to be consistently detected when changes occur in the scene[20]. Descriptors are specific codes generated based on the contextual environment (images within a neighborhood window) of the key points for key point discrimination. Without key points features, key points cannot be recognized.

The significant progress in this field has also set off a research boom in deep learning because deep neural network models trained using deep learning theory and large datasets have the strong expressive power to extract more abstract and essential features from the original input data, thus facilitating the solution of problems in areas such as image recognition[21].

Research[22] on key point detection methods has focused on finding more unique and discriminative image locations. Machine learning-based key point detection methods started early and have been introduced for different purposes in different methods, thus giving rise to various detection strategies. Immediately after that, some researchers proposed machine learning-based edge detection methods that were successfully applied. A robust key point detection algorithm based on segmented linear convolution was presented. The detection results showed that the learning-based key point detection method could achieve better results than the previous detection methods and has good invariance to light and seasonal changes[23]. Recently, another researcher[24] proposed a method combining key point detection and feature descriptors and introduced a multi-scale convolutional structure into the detection model. The feature descriptor aims to provide a discriminative representation of the target image block, which should be robust to environmental changes such as viewpoint or illumination. Although there have been many successful non-machine learning methods in this area, these are still hand-designed descriptors. Their effectiveness has been surpassed by new learning-based methods[25]. There are many types of these learning-based feature extraction methods, including unsupervised learning methods represented by self-encoders, supervised learning methods based on linear discriminant analysis, genetic algorithms, and convex optimization.

Based on this, some researchers[26,27] discussed and proposed a convolutional network with more image-centric regions and got better performance. Immediately after, a researcher's approach[28] uses a similar architecture and achieves the best results in a narrow-baseline stereo-matching task. Meanwhile, another researcher[29] proposed a compact descriptor extraction method to determine similarity based on Euclidean distance, using hard-to-score sample mining to improve the model's learning ability.

Current Status of Research on Anomalous Behavior Detection Based on Deep Learning

It is almost impossible to draw a boundary between abnormal events and normal events because the same behavior in different scenarios can be an abnormal event or a normal event; a car driving on the street is a normal event, while driving into a sidewalk is an abnormal event; a pedestrian running on a sports field is a normal event while changing the scenario to running outside a bank is an abnormal event[30]. In practical applications, the anomalies are unknown in advance, and it is impossible to learn models for all anomalous behaviors[31]. Usually, anomalous events are considered unexpected events which occur less frequently than normal events.

Anomalous behavior detection based on traditional machine learning generally uses manual feature methods to extract behavioral features. The key steps of abnormal behavior detection are feature extraction and abnormality detection, and the result of feature extraction determines the accuracy of abnormal behavior detection. Most previous anomalous behavior detection uses manual feature extraction methods to extract pedestrian appearance, motion, spatial, and temporal features to achieve anomalous behavior detection[32]. Aceto et al[33] analyzed the predictive ability of improved frailty index on pulmonary complications after major abdominal surgery in the elderly. The study compared and evaluated the respiratory risk of surgical patients in Catalonia and the predictive ability of the American sociophysical state classification. Gradient histogram as a statistical tool is often used to describe appearance features. While motion features are currently commonly used to extract optical flow features using optical flow histograms. With further research, spatial and temporal gradients and video spatial and temporal blocks have also been proposed to consider appearance and motion features[34]. The above manual features are often insufficient to characterize the target, so they are often not recommended for video surveillance scenarios with complex scenes.

Deep learning-based abnormal behavior detection is usually divided into three steps: 1) feature extraction, 2) construction of a normal behavior model, and 3) identification of abnormal behavior. In the training phase, feature extraction can be constructed manually or by deep learning methods to learn the features of normal behaviors in the original video clips and train the normal behavior model[35]. In the testing phase, feature information is extracted first, and the trained normal behavior model is used to classify it to determine whether it is an abnormal behavior. Deep learning-based anomaly detection can be divided into three basic frameworks based on whether feature extraction and anomaly detection are separated: feature extraction based on deep learning, learning based on normal features, and end-to-end anomaly metric learning[36]. The classification of deep learning-based anomalous behavior detection is shown in Figure 2.

In end-to-end anomaly metric-based learning, feature extraction and anomaly detection are fully unified, and the method uses neural networks to directly calculate anomaly scores. The raw data is used as input to directly learn and output anomaly scores to detect anomalous behavior[37]. Then, each video is represented as a packet, and each video clip is an instance in the packet; next, the C3D features in the video clip are extracted; finally, a fully connected neural network is trained using a new ranking loss function that calculates the highest scoring instance in the packet to locate anomalies.

Currently, only some articles[38,39] studied end-to-end anomaly metric-based learning methods, which have been proposed only in the last two years. End-to-end detection uses an unsupervised method that can identify anomalous frames from a large unlabeled number of videos, solving the problem of high data labeling cost and obtaining anomaly scores in just one step[40]. However, com-

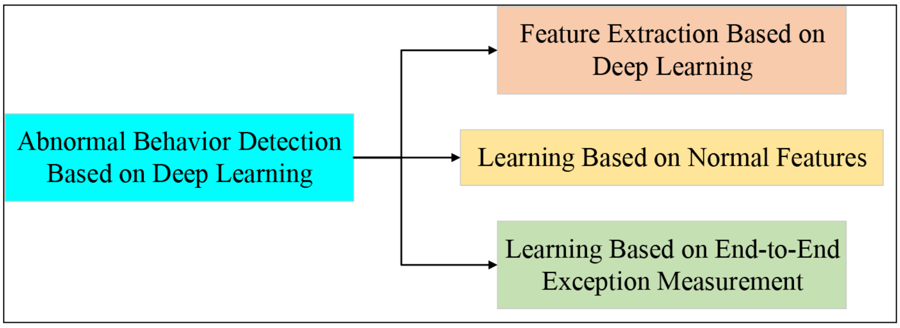
Abnormal Behavior Detection Based on Deep Learning
Feature Extraction Based on Deep Learning
Learning Based on Normal Features
Learning Based on End-to-End Exception Measurement

Figure 2. Classification of abnormal behavior detection based on deep learning.

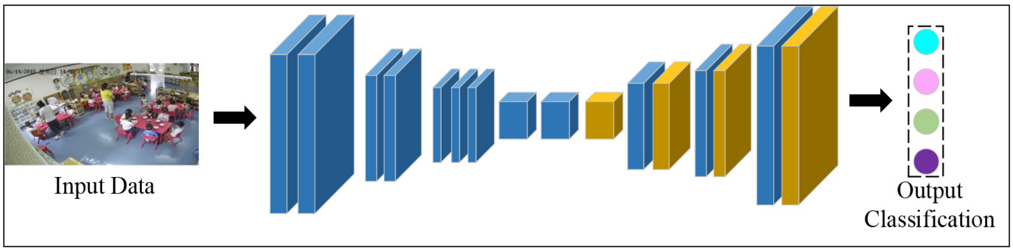

Figure 3. MotionNet network structure.

pared with previous supervised methods, there is still more room for performance improvement, and it is worthwhile for researchers to continue their in-depth research. In practical applications, the development of human anomaly detection technology is essential. In recent years, algorithms that apply deep learning techniques to anomalous behavior detection tasks have also emerged[41]. Although end-to-end approaches have been proposed for anomalous behavior detection tasks in the last two years, the generalization is insufficient. The performance of such unsupervised methods still has a particular gap compared with supervised methods. In practical research, anomaly detection models can be trained by collecting videos of different scenes and generalizing the learned models to videos of new scenes, which requires the high computational power of the device. Therefore, researching more pervasive anomaly detection algorithms is still a daunting challenge.

Algorithm Design

Feature extraction network model

With in-depth research on deep learning, video behavior classification technology has been developed rapidly. Traditional video behavior classification methods such as SIFT (Scale-invariant feature transform), HOG (Histogramof oriented gradients), and others require specialized knowledge and have huge limitations. As the speed and accuracy of neural networks gradually outperformed traditional algorithms, neural network models began to replace conventional algorithms as the foundation of the video behavior classification field[42]. However, neural network models still have huge shortcomings in capturing motion information. Then, researchers[43,44] started to use optical flow to help neural networks extract motion information, but optical flow can limit the speed of the models while bringing improvements in correctness. To address these problems, a neural network model is used to compute optical flow to improve the speed, fuse the space-time features as new features, and use a new network structure to process optical flow to improve the accuracy and speed of video behavior classification.

MotionNet network model

To address the problem that Convolutional Neural Network (CNN) is ineffective in extracting motion information from continuous frames, optical flow instead of RGB frames is input into the network to help CNN extract motion information. Therefore, using optical flow as input is equivalent to filtering out the spatial features in the original RGB frames and using motion features directly to improve the CNN's ability to extract motion signs.

The structure of MotionNet is shown in Figure 3, which is a neural network for extracting optical flow in the Hidden Two Stream network model. The optical flow computation calculate the velocity image between two adjacent frames. MotionNet treats the optical flow computation problem as a picture reconstruction problem and constructs the optical flow image by extracting the features of two adjacent images.

MotionNet is a fully convolutional network composed of a part of the down-sampling network and a part of the up-sampling network. The down-sampling network is composed of a series of convolutional layers that extract features layer by layer, and the main task is to extract the image features of two adjacent frames. The up-sampling network comprises a series of de-convolutional layers to recover the corresponding optical flow image from the features extracted by the up-sampling network[45]. The loss function is the key to the neural network, and the goodness of the loss function often determines the final performance of this neural network. For MotionNet, to better learn the features of optical flow, a Structure Similarity Loss Function (SSIM) is used to help the network learn the structure between RGB frames straight with a loss function:

$$L_{ssim} = \frac{1}{N}\sum\left(1 - SSIM\left(I_l - I_l'\right)\right)(1)$$

where the SSIM function is:

$$SSIM(x,y) = l(x,y)^{\alpha} + \left(c(x,y)^{\beta} + s(x,y)\right)^{\gamma}(2)$$

where x and y are the comparison of two images, l(x, y) is the comparison of brightness, c(x, y) is the comparison of design contrast, and (x, y) is the comparison of structure.

Optical Flow Feature (OFF) Network Model

To address the problem of slow computation of traditional optical flow, MotionNet is used to compute directly in the network without computing the optical flow in advance, which can significantly reduce the computation time. Instead of using optical flow as input, the video RGB frames are directly used as input, and MotionNet is used to calculate the optical flow features. In the network here, 11 consecutive video RGB frames are input. MotionNet generates ten optical flow frames, each with optical flow maps in x and y directions, and there are 20 optical flow maps in total, i.e., 224 * 224 * 3 * 11 for input and 224 * 224 * 2 * 10 for input. Finally, the optical flow generated by MotionNet is input to the traditional neural network for further extraction of temporal features[46]. For its network features and the network features, the OFF network is introduced into the optical stream feature extraction network, as shown in Figure 4. So, it helps the network to extract temporal features more fully.

The OFF network is inspired by the conventional constant brightness in optical flow, calculated as follows.

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)(3)$$

The OFF network can be added to the traditional CNN for a more efficient extraction of temporal features. The inception network is selected as the main network. After each convolutional layer of inception, the features generated by the inception network are input to the OFF network for further feature extraction. Finally, all the features generated by the OFF network are down-sampled from different layers due to the different sizes of the layers and aggregated together for classification prediction[47]. Since adding the OFF network to inception will cause the whole network to be too complex, the network cannot be trained well with just the final Softmax loss function, so multiple loss functions are added at different locations to adjust the network. In the overall training, the backbone network is trained first. Then, the parameters of the backbone network are fixed before training the OFF sub-network. Finally, all the fine-tuning is done.

Two-channel feature fusion model

From the analysis of the previous two subsections, temporal and spatial features are independent and complementary. When classifying a video, using only spatial or temporal features cannot accurately classify the video and will generate ambiguity. In a traditional two-channel network, the spatial features only input RGB frames, while the temporal features only take the processed optical stream as input. In this way, only one feature exists in each channel, RGB frames contain only spatial features, and it is difficult to extract temporal features directly. In addition, optical streams leave only motion information without spatial features. Therefore, it is difficult to fuse spatial and temporal features simultaneously in previous models. However, since the input of the temporal channel is no longer an optical stream but continuous RGB frames, many spatial features will be generated in the process of optical stream computation by MotionNet so that the proposed model can fuse spatial and temporal features to improve the correct rate.

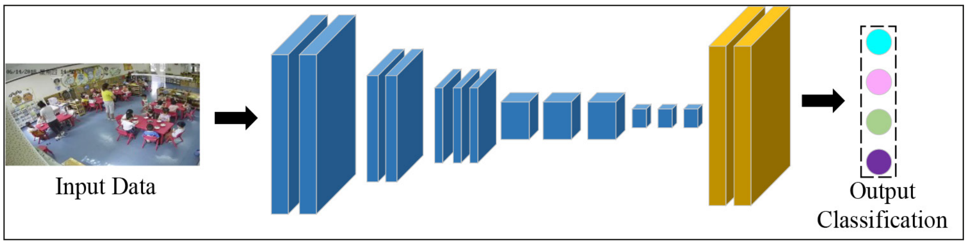

Figure 4. OFF network structure.

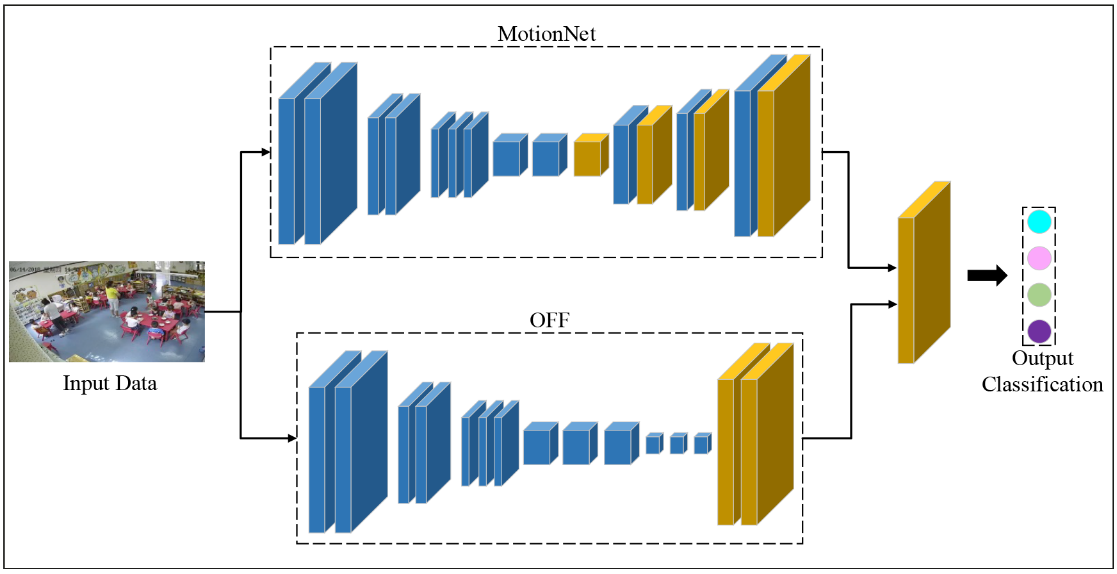

Figure 5. Two-channel network structure.

To classify video behavioral actions more accurately and effectively, this paper divides the video information into two main aspects: spatial and temporal features. Spatial features refer to the scene and item information contained in the RGB frames of a video. For a video, spatial features can be used to roughly classify its action categories. However, some actions are difficult to distinguish by spatial features alone, so it is necessary to use temporal features[48]. Temporal features refer to the movement of people or objects between two adjacent frames in a video, which mainly contains motion information. Similarly, it is also difficult to distinguish some actions by temporal features alone. The analysis should be combined with spatial features and temporal features to be able to extract and characterize the behaviors more comprehensively and to be able to classify the behaviors in the video more accurately.

In Figure 5, this paper adopts a two-channel network structure as shown in the figure. The network contains two independent single-channel models to extract spatial and temporal features. For the extraction of spatial features, RGB frames of the video are directly used as input, and a CNN is used to extract the corresponding spatial features; for temporal features, MotionNet is used to compute optical flow, and the input of MotionNet is also continuous video frames. Then, the computed optical flow is used as input. A CNN with an OFF network is used to further extract temporal features[49]. Finally, the extracted spatial and temporal features are weighted and fused to classify the final results.

The input to the network is a single frame of video RGB image because there may be different spatial features at different locations in the video. To address this factor, the video is divided into multiple segments equally, and the spatial features are extracted using the first frame for each segment separately. For example, if the playing basketball video is divided into three segments, one RGB image frame is extracted at the beginning and middle of the video and the end as input to extract spatial features. Finally, the spatial features of multiple videos are fused and classified.

Abnormal Behavior Detection Methods

Human behavior videos usually contain many complex backgrounds and camera shakes, and these complex situations can increase the difficulty of human behavior recognition. In this section, a cluster pooling method is proposed to eliminate the redundant information in the video. The aim of the section proposes a trajectory-weighted depth convolutional order pooling descriptor based on the proposed trajectory attention graph and cluster pooling method, combined with the order pooling method, which can effectively describe the dynamics of human behavior in long-time redundant videos in complex background environments.

Figure 6 shows the computational flow of the trajectory-weighted deep convolutional order pooling descriptor. The process mainly consists of the following five steps:

(1) All frames of the RGB video are input to the VGG-16 convolutional network to compute the convolutional feature maps, which are normalized using the spatial and temporal normalization method.

(2) Based on the RGB video, improved dense trajectories are computed, which can describe the trajectory of the moving target in the video.

(3) By weighting the trajectory attention map of each frame into the corresponding convolutional feature map, the trajectory-weighted convolutional features of the target region can be obtained for each frame.

(4) The clustering pooling method is used to reduce the redundant information in the time series of trajectory-weighted convolutional features.

(5) Order pooling is used to eliminate the redundant trajectory-weighted convolutional feature time series.

The attention map is widely used to localize target regions in the video, which is equal in length and width to the convolutional feature map. The convolutional feature map and the attention map in each frame correspond. The proposed trajectory attention graph is calculated by dense trajectory sub and improved dense trajectory sub, which can describe the trajectory of a person in a complex environment, so these trajectory points are continuously distributed in the motion target area of the video. The pixel value is determined by counting the number of trajectories in the perceptual field corresponding to each pixel point in the trajectory attention map. If the number of trajectories in the perceptual field increases, the corresponding pixel value will be larger. Therefore, the trajectory attention map reported here is more advantageous in character localization.

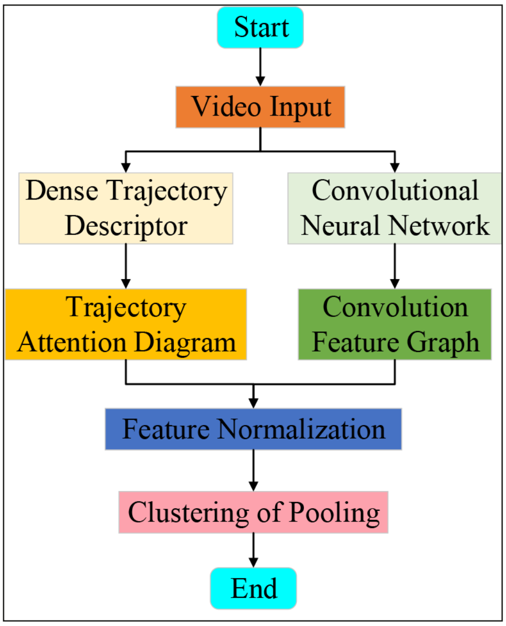
Start
Video Input
Dense Trajectory Descriptor
Convolutional Neural Network
Trajectory Attention Diagram
Convolution Feature Graph
Feature Normalization
Clustering of Pooling
End

Figure 6. Subcalculation flow depicted by sequence pooling.

The computation of dense trajectories consists of the following processes: first, in the initial video frame, the dense feature points are sampled every 5 pixels using dense grid sampling; then, the eigenvalues of the corresponding autocorrelation matrix are calculated for each dense feature point, and if the eigenvalues are low, it means that these points are in the gentle background area, and such points are removed by setting a threshold. Finally, for each feature point left behind, it is considered the starting point of the trajectory. The dense trajectory subsets can be obtained by tracking the trajectory starting points along the time, and its calculation process is as follows.

$$\left(x^{t+1}, y^{t+1}\right) = \left(x^{t}, y^{t}\right) + \left(M * \omega^{t}\right)\Big|_{\left(x^{t}, y^{t}\right)} \quad (4)$$

To calculate the convolutional features for each motion region in the video, the trajectory attention map is weighted into the convolutional feature map to obtain the trajectory-weighted convolutional features. The trajectory-weighted convolutional feature in the frame is:

$$U^{t} = \forall_{c \in C}\left(\sum\nolimits_{i,j}\left(a_{ij}^{t} \cdot f_{ij}^{t}\right)\right) (5)$$

Videos usually contain a large amount of redundant information, which can significantly increase the difficulty of coding the dynamics of human behavior. Therefore, this paper proposes a new clustering pooling method, which can effectively eliminate the redundant information in the video. Cluster pooling is used for video frame-level feature sequences, which cluster the sequences along the time dimension to eliminate redundancy. The clustering pooling first clusters the time sequences into multiple disjoint sequence segments. Then, it computes each segment's mean vector as the corresponding segment's representation.

$$\begin{cases} m^{i} = \dfrac{1}{i}\sum\limits_{\tau=1}^{i} d^{\tau} \\ d = D\left(\mu^{1}, U^{2}\right) \end{cases} \quad (6)$$

Finally, the trajectory-weighted deep convolutional order pooling descriptor is obtained by introducing the order pooling method to encode the dynamic information in the trajectory-weighted convolutional sequences.

Results

Experimental Data Set and Pre-Processing

The SDUFall, Weizmann, and HMDB51 datasets, widely used in human behavior recognition research, are used in the experiment. The SDUFall dataset is captured by a 1.5 m high Kinect camera, which contains six types of behaviors: falling, bending, squatting, sitting, lying down, and walking; the Weizmann dataset contains ten basic daily behaviors such as bending, jumping, running and walking, which are captured by ten volunteers, and each type of behavior contains nine video frame sequences. The HMDB51 data contains 7,000 video samples and 51 behavioral categories, each with at least 101 video samples. All the data are obtained from videos and movies on the web, so the dataset is very challenging.

Training of spatial flow model

To make the model converge faster and get a better classification rate during training, the inception model is first trained using the ImageNet dataset. Next, the trained model parameters are copied to the spatial flow model as the initial parameters. Then, the training set of the corresponding dataset is used to further adjust the model parameters and learn the correct features. Finally, the score results of spatial stream classification are stored.

Training of the time-streaming model

Since the time-streaming model network is complex, to adequately train the whole model, a step-by-step training method is used in this part of the model training, i.e., the backbone network is trained first. Then, the sub-networks are trained, and the overall network is adjusted.

Results of Feature Extraction Experiments

In the feature extraction experiments, the MotionNet and OFF network models were used to extract features and compared with our proposed two-channel feature extraction method. The results of the prediction accuracy of the three network models with the training data are shown in Figure 7, from which it can be seen that the two-channel model has the highest accuracy of the three data sets. Specifically, the two-channel model has 39 categories higher than MotionNet in SDUFall, 49 categories with the same correct rate as MotionNet, and the correct rate of most categories is 85% higher. For the ApplyEyeMakeup category, MotionNet is only 68%, while the proposed model can reach 84%. Similarly, in the category of Lunges, MotionNet is only 72%, while this model can achieve 89%. Compared with MotionNet, the proposed model adds spatial and temporal features and an OFF network to the CNN dealing with optical flow features, which can improve the correct rate significantly. However, in the three actions of brushing teeth, hammering, and nunchucks, the correct rate of this model is less than 60% because these actions are single and easily confused with other actions, resulting in a low correct classification rate.

Then, the running speed of each model is compared and, for better comparison, the feature extraction method using only optical flow is also added. The speed of the models is shown in Figure 8. This is because both the MotionNet model and the model proposed here use RGB images as input. In general, the model using optical flow has a significant advantage in speed, but it has a low correctness rate due to the lack of temporal features as input. The optical flow features can help the model improve the correct rate significantly for the optical flow model. Still, because the traditional optical flow algorithm is slow, this model uses the neural network MotionNet to replace the traditional optical flow extraction, which can improve the speed significantly and ensure the correct rate of the model is higher than most models.

Finally, the comparison is performed on the HMDB51 dataset. Since the HMDB51 dataset is more complex and has more diverse actions compared to the SDUFall scene, most of the models have a lower correct rate in terms of correctness. Experimental results on HMDB51 dataset are revealed in Table I.

From the numerical results in Table I, the accuracy rate of this model on this dataset is still slightly higher than that of other models when only video RGB frames are used as input. Compared to OFF, it can reach 88.64% on SDUFall and only 63.81% on HMDB51. In comparison, the proposed model can reach 72.09%, because the HMDB51 dataset is complex, and motion information cannot be directly obtained using RGB frames. The proposed model uses MotionNet to

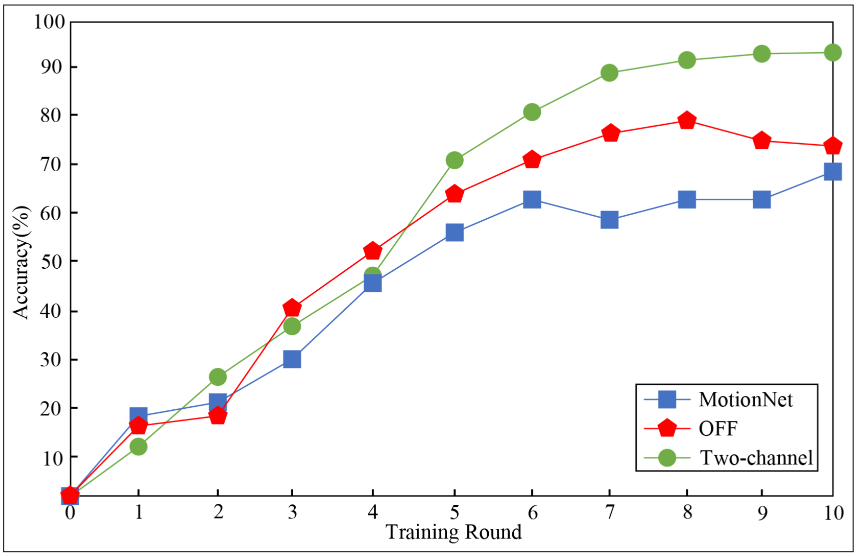

Figure 7. Changes in model prediction accuracy with training data.

calculate optical flow on RGB frames, which has certain advantages, so the accuracy rate is slightly higher than that of MotionNet.

Experimental Results of Anomalous Behavior Detection

An important parameter in clustering pooling is the threshold value, which determines the length and number of clustered sequence segments. First, the time series length reduction percentage in the SDUFall dataset is counted as the threshold value shifts. Then, the recognition accuracy on the SDUFall dataset was tested when the threshold value was taken in the range of 0-1, so the optimal threshold value could be selected. The results are shown in Figure 9. From Figure 9, the larger the threshold value of cluster pooling, the shorter the time series length after removing redundancy, so the cluster pooling method can effectively reduce the redundancy in the time series. Figure 9 shows that the accuracy rate can reach more than 95% when the threshold value is 0.8, so the threshold value will be set to 0.8 in all future experiments.

A related comparison experiment is set to further evaluate our trajectory attention map. In the comparison experiments, the trajectory attention map is removed to use three other features instead of the trajectory-weighted convolutional features. The first comparison feature is the maximum pooling feature, which results from the maximum direct pooling of the convolutional feature map. The second comparison is the average pooling feature, resulting from the convolutional feature map's direct global average pooling. The third feature is the fully connected feature. These three contrast features do not consider the location of the moving characters in the video, so they cannot encode the visual features of the character regions in a targeted way. Then, these three contrast features are pooled by clustering and pooled by order to obtain two contrast descriptors, which are called maximum convolutional order pooling descriptor (MDRD), average convolutional order

Table I. Experimental results on the HMDB51 dataset.

	Data Set		
Method	**SDUFall**	**Weizmann**	**HMDB51**
MotionNet	82.57%	86.39%	63.81%
OFF	78.73%	81.66%	60.75%
Two-channel	88.64%	89.28%	72.09%

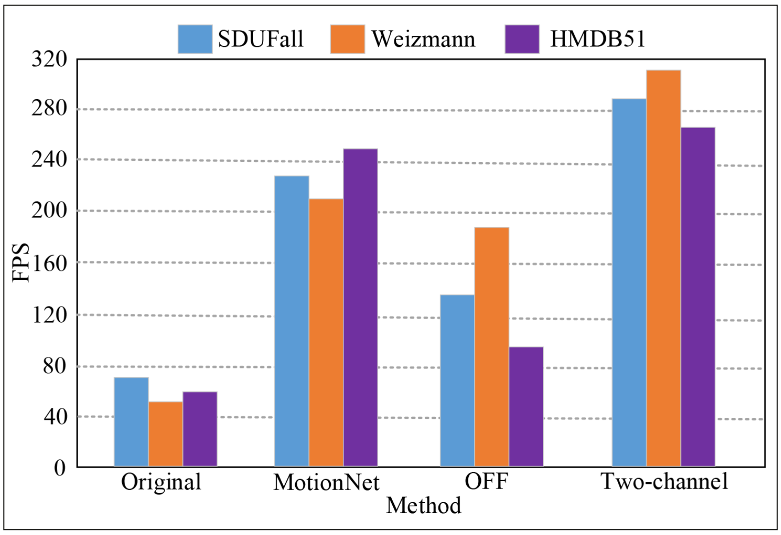

Figure 8. Experimental results of model running speed.

pooling descriptor (ADRD), and fully connected order pooling descriptor (FDRD), respectively.

The results of the three comparison descriptors with the trajectory-weighted deep convolutional order pooling descriptor on the dataset are shown in Table II.

As Table II suggests, the TDRD (Tone-Mapped Dynamic Range Descriptor) descriptor obtains the best results of 92.57%, which is 3.64%, 2.58%, and 1.73% higher than the three comparison descriptors MDRD, ADRD, and HDRD (High Dynamic Range Descriptor), respectively. The main reason is that in human behavior recognition, all three contrast descriptors do not consider the position of the person area in the video, so they cannot encode the visual features of the person area. The trajectory-weighted depth convolutional sequence pool descriptor can efficiently locate the area of people in the video and encode visual features, so the descriptor is more efficient than these three contrast descriptors. This finding suggests that the use of trajectory-weighted deep convolutional sequence pool descriptors may be more efficient at extracting and detecting behavioral traits in children with autism spectrum disorder than other commonly used descriptors.

Discussion

Childhood autism spectrum disorder is a multifactorial, widespread psychological developmental disorder with a prevalence of 1% to 2%. This disorder has a high disability rate and imposes a huge burden on families and society economically. Therefore, it is crucial to have an active and effective intervention treatment. Based on the above background, this paper proposes a deep learning-based visual strategy intervention method for autism spectrum disorders, combined with artificial intelligence technology to assist in treating children with autism spectrum disorders. First, to better extract children's behavioral features, the convolutional neural network MotionNet is used to compute the optical flow us-

Table II. Descriptors used to compare experimental results.

	Data Set		
Description Subtype	**SDUFall**	**Weizmann**	**HMDB51**
MDRD	88.57%	90.04%	86.23%
ADRD	89.34%	91.57%	87.62%
FDRD	89.81%	92.37%	89.45%
Ours (TDRD)	92.85%	93.64%	92.28%

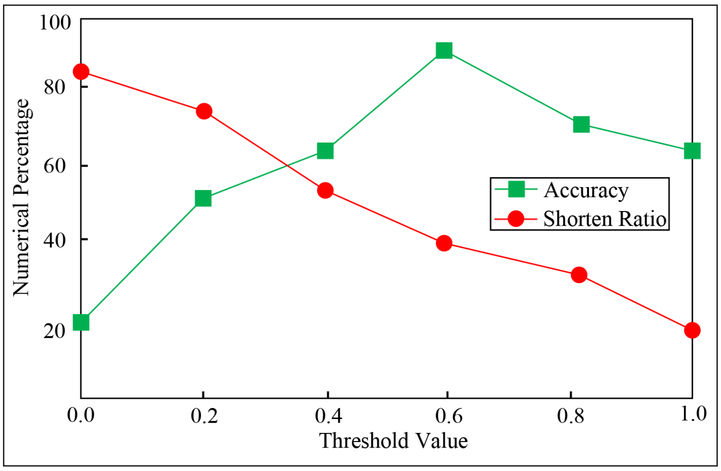

Figure 9. Experimental results of clustering pooling threshold.

ing the information between two adjacent frames, which significantly improves the network's speed. Meanwhile, to address the problem of insufficient optical flow feature extraction, the OFF network is proposed to be added to the traditional network of optical flow extraction to fully extract temporal features in optical flow. Then, regarding the drawback of the many redundant frames contained in video, a method of pooling descriptors based on trajectory-weighted deep convolutional order is proposed. The descriptor is calculated based on the trajectory attention map, convolutional feature map, clustering pooling method, and order pooling method, which can effectively describe the behavioral dynamics of people in complex background environments in long-time redundant videos. Finally, the network of this paper is evaluated for accuracy on three datasets. The results show that our model is more accurate and faster than the traditional model. Using this method, sick children can be monitored in real-time and intervene with visual strategies to guide them to better integrate into the group.

Conclusions

This method and visual intervention for children with autism spectrum disorders can help them to overcome social barriers. However, there are some shortcomings. First, this paper only uses video as a data source and does not use other types of data. Second, only three datasets are used for evaluation, so testing on more datasets is needed to verify the generalization performance of the model. In addition, the individual differences of children are not taken into account. Future research needs to pay more attention to the individual differences of children and develop more personalized interventions. In future studies, data sources can be expanded to use multiple types of data to improve the robustness and generalization performance of the model. More personalized intervention programmes should be developed. It is necessary to consider the individual differences of children, combine psychology, education, and other disciplines to develop more effective intervention strategies, further improve the speed and accuracy of the model, and achieve more real-time intervention. These research directions will help to further improve the treatment effect of autism spectrum disorder in children, reduce the burden on children and families, and are also of great significance to the development of social health.

Conflict of Interest

The Authors declare that they have no conflict of interests.

Availability of Data and Materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Funding

None.

Ethics Approval

Not applicable.

Informed Consent

Not applicable.

ORCID ID

Y. Lin: 0009-0008-6009-3495.

References

1) Ganz JB, Lashley E, Rispoli MJ. Non-responsiveness to intervention: children with autism spectrum disorders who do not rapidly respond to communication interventions. Dev Neurorehabil 2010; 13: 399-407.
2) Liao ST, Hwang YS, Chen YJ, Lee P, Chen SJ, Lin LY. Home-based DIR/Floortime intervention program for preschool children with autism spectrum disorders: preliminary findings. Phys Occup Ther Pediatr 2014; 34: 356-367.
3) Kasari C, Smith T. Interventions in schools for children with autism spectrum disorder: methods and recommendations. Autism 2013; 17: 254-267.
4) Girolametto L, Sussman F, Weitzman E. Using case study methods to investigate the effects of interactive intervention for children with autism spectrum disorders. J Commun Disord 2007; 40: 470-492.
5) Vandermeer J, Beamish W, Milford T, Lang W. iPad-presented social stories for young children with autism. Dev Neurorehabil 2015; 18: 75-81.
6) Washington P, Park N, Srivastava P, Voss C, Kline A, Varma M, Tariq Q, Kalantarian H, Schwartz J, Patnaik R, Chrisman B, Stockham N, Paskov K, Haber N, Wall DP. Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging 2020; 5: 759-769.
7) Jaliaawala MS, Khan RA. Can autism be catered with artificial intelligence-assisted intervention technology? A comprehensive survey. Artificial Intelligence Review 2020; 53: 1039-1069.
8) Anagnostopoulou P, Alexandropoulou V, Lorentzou G, Lykothanasi A, Ntaountaki P, Drigas A. Artificial intelligence in autism assessment. International Journal of Emerging Technologies in Learning (iJET), 2020; 15: 95-107.
9) Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019; 21: e13216.
10) Myers SM, Johnson CP. Management of children with autism spectrum disorders. Pediatrics 2007; 120: 1162-1182.
11) Cao XJ, Liu XQ. Artificial intelligence-assisted psychosis risk screening in adolescents: Practices and challenges. World J Psychiatry 2022; 12: 1287-1297.
12) Wang Y, Wen Q, Jin L, Chen W. Artificial Intelligence-Assisted Renal Pathology: Advances and Prospects. J Clin Med 2022; 11: 4918.
13) Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls CJ, Sohl K, Rouhbakhsh R, Narasimhan A, Romain J, Golla S, Shareef S, Ostrovsky A, Shannon J, Kraft C, Liu-Mayo S, Abbas H, Gal-Szabo DE, Wall DP, Taraman S. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. NPJ Digit Med 2022; 5: 57.
14) Marciano F, Venutolo G, Ingenito CM, Verbeni A, Terracciano C, Plunk E, Garaci F, Cavallo A, Fasano A. Artificial Intelligence: the “Trait D’Union” in Different Analysis Approaches of Autism Spectrum Disorder Studies. Curr Med Chem 2021; 28: 6591-6618.
15) Rapakoulia T, Theofilatos K, Kleftogiannis D, Likothanasis S, Tsakalidis A, Mavroudi S. EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms. Bioinformatics 2014; 30: 2324-2333.
16) Rahman MM, Usman OL, Muniyandi RC, Sahran S, Mohamed S, Razak RA. A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder. Brain Sci 2020; 10: 949.
17) Virnes M, Kärnä E, Vellonen V. Review of research on children with autism spectrum disorder and the use of technology. Journal of Special Education Technology 2015; 30: 13-27.
18) Bharathi G, Jayaramayya K, Balasubramanian V, Vellingiri B. The potential role of rhythmic entrainment and music therapy intervention for individuals with autism spectrum disorders. J Exerc Rehabil 2019; 15: 180-186.
19) Si T, Zhu Y, Zongni L. Application and Prospect of Immersive Virtual Reality Technology in Rehabilitation Practice of Autistic Children. Applied & Educational Psychology 2022; 3: 59-67.
20) Bahado-Singh RO, Vishweswaraiah S, Aydas B, Radhakrishna U. Artificial intelligence and placental DNA methylation: newborn prediction and molecular mechanisms of autism in preterm children. J Matern Fetal Neonatal Med 2022; 35: 8150-8159.
21) Liang H, Sun X, Sun Y, Gao Y. Text feature extraction based on deep learning: a review. EURASIP J Wirel Commun Netw 2017; 2017: 211.
22) Mohan A, Poobal S. Crack detection using image processing: A critical review and analysis. Alexandria Engineering Journal 2018; 57: 787-798.
23) Zeng W, Gautam A, Huson DH. DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome. Bioinformatics 2022; 38: 4670-4676.
24) Engel J, Koltun V, Cremers D. Direct Sparse Odometry. IEEE Trans Pattern Anal Mach Intell 2018; 40: 611-625.

25) Hu Z, Yang Z, Lafata KJ, Yin FF, Wang C. A radiomics-boosted deep-learning model for COVID-19 and non-COVID-19 pneumonia classification using chest x-ray images. Med Phys 2022; 49: 3213-3222.

26) Eulenberg P, K?hler N, Blasi T, Filby A, Carpenter AE, Rees P, Theis FJ, Wolf FA. Reconstructing cell cycle and disease progression using deep learning. Nat Commun 2017; 8: 463.

27) Yu H, Song K, Meng Q, Yan Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Transactions on Instrumentation and Measurement, 2019; 69: 1493-1504.

28) Rogge S, Schiopu I, Munteanu A. Depth estimation for light-field images using stereo matching and convolutional neural networks. Sensors 2020; 20: 6188.

29) Dube R, Cramariuc A, Dugas D, et al. SegMap: Segment-based mapping and localization using data-driven descriptors. The International Journal of Robotics Research 2020; 39: 339-355.

30) Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, Kudo M, Haida K, Kuroda J, Yanagiya R, Saitoh E, Hoshinaga K, Yuzawa Y, Suzuki A. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep 2019; 9: 11862.

31) Abosaq HA, Ramzan M, Althobiani F, Abid A, Aamir KM, Abdushkour H, Irfan M, Gommosani ME, Ghonaim SM, Shamji VR, Rahman S. Unusual Driver Behavior Detection in Videos Using Deep Learning Models. Sensors (Basel) 2022; 23: 311.

32) Parastarfeizabadi M, Sillitoe RV, Kouzani AZ. Multi-disease Deep Brain Stimulation. IEEE Access 2020; 8: 216933-216947.

33) Aceto P, Perilli V, Luca E, Schipa C, Calabrese C, Fortunato G, Marusco I, Lai C, Sollazzi L. Predictive power of modified frailty index score for pulmonary complications after major abdominal surgery in the elderly: a single centre prospective cohort study. Eur Rev Med Pharmacol Sci 2021; 25: 3798-3802.

34) Ogallo W, Tadesse GA, Speakman S, Walcott-Bryant A. Detection of Anomalous Patterns Associated with the Impact of Medications on 30-Day Hospital Readmission Rates in Diabetes Care. AMIA Jt Summits Transl Sci Proc 2021; 2021: 495-504.

35) Lian B, Kartal Y, Lewis FL, Mikulski DG, Hudas GR, Wan Y, Davoudi A. Anomaly Detection and Correction of Optimizing Autonomous Systems With Inverse Reinforcement Learning. IEEE Trans Cybern 2022; PP.

36) Knights J, Heidary Z, Cochran JM. Detection of Behavioral Anomalies in Medication Adherence Patterns Among Patients With Serious Mental Illness Engaged With a Digital Medicine System. JMIR Ment Health 2020; 7: e21378.

37) Parastarfeizabadi M, Sillitoe RV, Kouzani AZ. Multi-disease Deep Brain Stimulation. IEEE Access 2020; 8: 216933-216947.

38) Ibidunmoye O, Rezaie AR, Elmroth E. Adaptive anomaly detection in performance metric streams. IEEE Transactions on Network and Service Management 2017; 15: 217-231.

39) Takimoto H, Seki J, F. Situju S, Kanagawa A. Anomaly detection using siamese network with attention mechanism for few-shot learning. Applied Artificial Intelligence 2022; 36: 2094885.

40) Aldayri A, Albattah W. Taxonomy of Anomaly Detection Techniques in Crowd Scenes. Sensors (Basel) 2022; 22: 6080.

41) Homayouni H, Ray I, Ghosh S, Gondalia S, Kahn MG. Anomaly Detection in COVID-19 Time-Series Data. SN Comput Sci 2021; 2: 279.

42) Kim H, Shon T. Industrial network-based behavioral anomaly detection in AI-enabled smart manufacturing. J Supercomput 2022; 78: 13554-13563.

43) Lv D, Luktarhan N, Chen Y. ConAnomaly: Content-Based Anomaly Detection for System Logs. Sensors (Basel) 2021; 21: 6125.

44) Liu X, Li X, Shi Q, Xu C, Tang Y. UAV attitude estimation based on MARG and optical flow sensors using gated recurrent unit. International Journal of Distributed Sensor Networks 2021; 17: 15501477211009814.

45) Limon-Cantu D, Alarcon-Aquino V. Multiresolution dendritic cell algorithm for network anomaly detection. PeerJ Comput Sci 2021; 7: e749.

46) Hirayama M, Kawato M, Jordan MI. The Cascade Neural Network Model and a Speed-Accuracy Trade-Off of Arm Movement. J Mot Behav 1993; 25: 162-174.

47) Renganathan V. Overview of artificial neural network models in the biomedical domain. Bratisl Lek Listy 2019; 120: 536-540.

48) Wang J, Liu S. Visual Information Computing and Processing Model Based on Artificial Neural Network. Comput Intell Neurosci 2022; 2022: 4713311.

49) Dong H, Wang X. Identification of Signature Genes and Construction of an Artificial Neural Network Model of Prostate Cancer. J Healthc Eng 2022; 2022: 1562511.