Multimedia Systems (2025) 31:312
https://doi.org/10.1007/s00530-025-01885-4

REGULAR PAPER

ConASD: Contrastive Few Shot Learning for Detecting Autism Spectrum Disorder via Eye Tracking Scanpath

Sharifah Mousli[1,3] · Sona Taheri[1] · Estrid He[2]

Received: 14 February 2025 / Accepted: 2 June 2025 / Published online: 5 August 2025

Abstract

Detecting Autism Spectrum Disorder (ASD) using Eye Tracking (ET) datasets is a challenging task and has been a long-standing problem. Recently, there has been a trend of developing ASD diagnosis models based on machine learning (ML), especially deep learning techniques. In this paper, we show that these existing methods still struggle to make accurate diagnoses in few-shot learning (FSL) settings, where the data available for training is limited in amount and imbalanced in nature. To address this challenge, we propose a model, named ConASD, for effective diagnosis of ASD under the FSL setting. The proposed model is a two-stage framework: it first trains an encoder for ET images using supervised contrastive learning, followed by fine-tuning a classifier for final diagnosis. With the contrastive learning strategy, the pre-trained encoder can better capture the discriminative features of the eye-tracking images, even with limited training data, and ultimately leads to better diagnosis accuracy and better generalization to unseen data. We evaluate the proposed ConASD model using two real-world ET datasets. The results demonstrate that ConASD outperforms existing approaches, particularly in few-shot scenarios, by up-to 7% improvement in terms of F1 scores. The results in this paper highlight the potential of using contrastive learning as a powerful tool, particularly in real-world medical scenarios where class imbalance is frequent and the data is limited.

Keywords Autism spectrum disorder · Eye tracking scanpath · Few shot learning · Contrastive learning

1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting brain function, which causes differences in social communication and interaction. Studies [1] show that early detection of ASD can significantly benefit children and families with improved outcomes in language skills, adaptive behaviour, and educational support. However, the process of diagnosing ASD continues to be a difficult task that requires the use of cognitive assessments and comprehensive clinical evaluations.

Recently, computer-aided technologies started to play a crucial role in assisting ASD diagnosis, including magnetic resonance imaging, electroencephalography, and Eye Tracking (ET) [2]. In this paper, we focus on the ET technique, a non-invasive method that involves capturing, tracking, and measuring the movements of the eyes or the exact focus point, known as the point of gaze, where a person's eyes are fixed within a visual scene [3].

Understanding and making decisions based on ET scanpaths is non-trivial, requiring extensive expertise of domain knowledge. Figure 1 presents some example scanpaths of Typically Developing (TD) children and children with ASD symptoms. Scanpaths of children with ASD are generally less focused, with randomness in eye movements and less fixation on focus points. However, measuring the fixation of eye movements from scanpaths is challenging as

Communicated by Bing-kun Bao.

Sona Taheri and Estrid He have contributed equally to this work.

✉ Sharifah Mousli
s3635335@student.rmit.edu.au

Sona Taheri
sona.taheri@rmit.edu.au

Estrid He
estrid.he@rmit.edu.au

1 School of Mathematical Sciences, RMIT University, Swanston Street, Melbourne, VIC 3000, Australia

2 School of Computing Technologies, RMIT University, Swanston Street, Melbourne, VIC 3000, Australia

3 Faculty of Science, University of Jeddah, Setten Road, Jeddah 23445, Makkah, Saudi Arabia

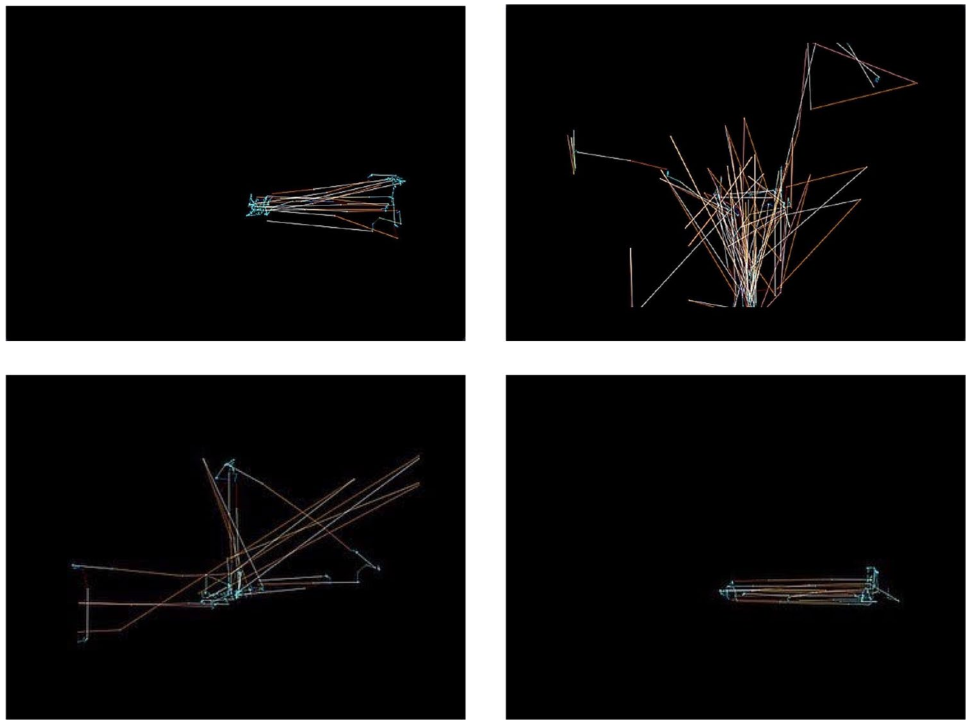


Fig. 1 Vizualization of scanpaths of TD and ASD participants. Top: easy classification case; Bottom: hard classification case. Left: participants without ASD (i.e., TD); Right: participants diagnosed with ASD (sourced from [16])

participants' focus points can vary drastically due to their own interests and experiment contexts, as shown in Fig. 1. Many studies seek machine learning (ML) aided techniques to extract useful features from scanpaths for assistive ASD diagnosis. For instance, in [4–7], ET data from both ASD and TD individuals is labeled by domain experts. Then a classifier is trained to identify new data accurately. Unsupervised learning techniques have been explored to identify ASD without the requirement for labeled data. They discern patterns or clusters within the ET data autonomously, uncovering hidden structures, such as ASD subtypes, that may not be readily apparent through conventional supervised ML methods [8–11].

Despite significant research effort in leveraging ML models for detecting ASD through ET techniques, there are still remaining limitations [4, 12, 13]. First, data-driven techniques such as ML models, especially DL models usually require a large-scale of high-quality training data. This is not readily available in ASD research, as data collection faces privacy issues [4, 12–15]. Thus, there has been increasing interest in tackling the scenario of Few Shot Learning (FSL), which aims at enhancing model performance when training data is extremely limited. Second, achieving high detection accuracy is challenging with ET scanpaths, as scanpaths of ASD children have only minor visual differences compared to those of TD children, as shown in Fig. 1.

Most existing works on analyzing ET scanpaths for ASD diagnosis mainly formulate the problem as an image classification task. A deep learning (DL) model, e.g., convolutional neural networks, takes images of scanpaths as input and learns to label the image into ASD or TD [17, 18]. Although promising results are obtained, the scenario of FSL is overlooked by literature. This, however, is crucial for ASD research, where high-quality training data is expensive to obtain.

This paper addresses the challenge of data limitation in model development for ET-based ASD diagnosis. In particular, we focus on the loss function for model training. We observe that the traditional Cross-Entropy (CE) loss for training a classifier is suboptimal for FSL as it is accuracy-oriented [19]. In ASD ET images, when the data available for training is limited, CE cannot guide a model to encode instances into tight and well-separated clusters. Inspired by the recent success of contrastive learning, we incorporate contrastive loss in model training [20, 21]. With this, the model is guided to refine feature embeddings by attracting similar (intra-class) pairs and repelling dissimilar (inter-class) pairs, effectively addressing the issue of CE loss [22]. We conduct extensive experiments on two real-world datasets. The results show that the incorporation of contrastive learning leads to significant improvements in predicting accuracy, especially when the number of training samples is limited. We also visualize the learned latent embeddings within the model, which also demonstrates the high-quality of the clusters formed by contrastive learning.

The rest of this paper is structured as follows. Section 2 discusses existing DL models in ASD detection. In Sect. 3, we present our proposed method for ASD detection.

Section 4 presents the experimental results with discussions on our empirical findings. Section 5 concludes this paper.

2 Related works

This section reviews related works from three perspectives: (1) ML based eye tracking for ASD diagnosis, (2) Contrastive learning applied to ASD diagnosis, and (3) FSL in ASD diagnosis.

2.1 ML based eye tracking for ASD diagnosis

Diagnosing ASD can be challenging, as there are no definitive pathological or radiological examination tools. Early signs of ASD may include behaviors such as avoiding eye contact and progressively develop more observable signs such as diminished focus on social cues, distinctive eye movement, and shorter fixation duration [23, 24]. Thus, researchers are exploring these eye movement anomalies to develop a theoretical framework for ASD screening through ET [5, 12, 13, 25], and use ET as an effective screening tool during the diagnosis.

There exist several studies that incorporate ML models to analyze scanpaths from eye tracking for assistive ASD diagnosis [26]. Elbattah et al. [13] aim at feature extraction over scanpaths to differentiate children with ASD with various severity of symptoms. They present a comparative analysis of four feature extraction techniques, namely principal component analysis, t-distributed stochastic neighbour embedding (t-SNE), and auto-encoders, with results showing the superiority of auto-encoders. In addition to unsupervised learning, several other studies formulate the problem as a classification task, i.e., determining if scanpaths are sampled from ASD or TD children. For instance, Ahuja et al. [27] propose a pipeline for analyzing scanpaths of TD and ASD children as they watch four short video clips. Five features are extracted from each clip. The features extracted from different clips are integrated and then fed to a support vector machine model for classification.

Recent advancements in this domain have shown a shift from classical ML techniques to deep models. Carette et al. [15] show that artificial neural networks can achieve high classification accuracy in identifying ASD and TD children. Atyabi et al. [18] develop spatial eye and allow the differentiation of ET scanpaths images of individuals with ASD and TD individuals using convolutional neural networks. A comparative study further demonstrates that hybrid models combining DL and ML outperformed standalone ML and DL models. In Ahmed et al. [17], a hybrid model of Convolutional Neural Networks (CNNs) for deep feature map extraction and support vector machines for feature classification for ASD diagnosis were employed.

2.2 Contrastive learning for ASD diagnosis

Contrastive learning has gained wide success in various domains over the past few years, including computer vision [20, 21], natural language processing [28, 29], and many more. In particular, the Supervised Contrastive Learning (SCL), proposed by [21], extends the framework of self-CL to the setting of supervised learning. During training, a predictive model (usually consisting of an input encoder and a classification layer) is guided to pull the feature representation of a data instance in the latent space *close* to the representations of its positive instances (e.g., those instances from the same class) while pushing it away from its negative instances (e.g., those instances from other classes). Studies have shown that SCL generates more effective representations compared to the case where CE loss is solely for model training [21, 30]. There are limited works that incorporate contrastive learning for ASD detection. Rani et al. [31] adapt contrastive learning to extract discriminative features from action video clips recorded. In Auriau et al. [32], SCL loss is applied to structural Magnetic Resonance Imaging data to distinguish ASD from other mental disorders, showing that SCL loss effectively identified structural abnormalities and achieved superior performance compared to CE loss. However, there exist no studies that incorporate contrastive learning for analyzing ET scanpaths to assist ASD detection.

2.3 Few shot learning in ASD diagnosis

In medical domains such as ASD research, obtaining high-quality data at large scale is usually challenging due to privacy issues and the requirement of domain expertise in the process of data collection. This challenges data-driven techniques such as DL models, whose effectiveness is heavily reliant on the quantity and quality of their training data. Thus, investigation into the scenario of FSL, i.e., only limited data instances are available for training has attracted attention. The Zhang et al. work [33] studies the facial expressions in an hour-long video recording from an Autism Diagnostic Observation Schedule interview. Their proposed model addresses the challenge of FSL by employing the scene-level information fusion [33]. Similarly, Dhinagar et al. [34] propose a joint learning framework that integrates multiple sites using MRI scans collected from 38 imaging sites as a part of the Autism Brain Imaging Data Exchange. The results demonstrated that the model can be adapted to unseen sites by fine-tuning under an FSL setting.

The study by Wang et al. [35] introduced a few-shot learning framework that integrates a Siamese Neural Network with a wide and deep architecture for an ASD screening dataset. This model was compared with other baselines to demonstrate its ability to effectively capture both linear and nonlinear relationships from a limited number of ASD samples. Similarly, Fawzy et al. [36] developed a model combining MobileNetV2 with two-way shot learning to diagnose ASD using children's facial images. The training set included four samples each of ASD and TD children's faces. The results indicated improved model performance when using cosine similarity over Euclidean distance, as cosine similarity effectively manages the high-dimensional feature space of human faces [37].

A recent study by Bennett et al. [38] introduced the SMOTE_FSL algorithm for diagnosing ASD in few-shot learning settings. This algorithm integrates machine learning techniques to effectively learn from limited data. While the original SMOTE algorithm [39] is widely used to address class imbalance, this study developed a custom variation tailored for FSL scenarios. The proposed SMOTE_FSL was compared with state-of-the-art deep learning methods in similar FSL contexts, including Autoencoders (AE), Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs). All models were evaluated using two digital health sensor datasets collected via robots and wearable devices.

In this paper, we propose to use SCL to address the challenge of FSL in ASD detection. Our methods differ from the above methods, as, to the best of our knowledge, no prior study has applied SCL in an FSL setting to ASD ET scanpaths.

3 Methodology

In this study, we focus on the task of ASD detection using ET scanpaths. Our aim is to propose a DL model capable of effectively extracting discriminative features from ET scanpaths for ASD diagnosis. In particular, we address the challenge of limited data availability, a common issue in ASD research. To overcome this, we propose a framework based on contrastive learning, which optimizes the embeddings of input ET images within a deep latent space.

Figure 2 illustrates the overall framework proposed for ASD detection. The framework consists of two stages: (1) Training of the encoder; and (2) Training of the classifier. We describe each step in detail below.

Training of the encoder. Given a set of images of ET scanpaths $\mathcal{D}_k$ represents the few-shot training samples, where $k \leq N$ and N is the total number of scanpaths in the training set. $\mathcal{D}_k = \{\boldsymbol{x}_i, y_i\}_{i=1,\dots,N}$, where $\boldsymbol{x}_i$ represents the input features (e.g., pixel vectors) of an ET scanpath, y_i is the corresponding ground-truth label. This stage trains an encoder capable of extracting discriminative features from $\boldsymbol{x}_i$. Specifically, the encoder learns a mapping function that transforms $\boldsymbol{x}_i$ into $\boldsymbol{f}_i$, where $\boldsymbol{f}_i$ encodes the key features of $\boldsymbol{x}_i$, enabling it to be effectively distinguished from other samples with significantly different characteristics.

Inspired by [21], we leverage contrastive learning for training the encoder. Specifically, we employ ResNet-50 as the backbone network for encoding input images, denoted as $Enc(\cdot)$. ResNet-50 consists of 50 CNN layers with residual learning, designed to address the vanishing gradient problem. Given an input image $\boldsymbol{x}_i$, a random image augmentation technique-chosen from a predefined set of operations such as rotation, cropping, and flipping-is applied, resulting in an augmented version of the image $\boldsymbol{x}'_i$. Both the original image and its augmented images are then encoded using the backbone network (e.g., $\boldsymbol{f}_i = Enc(\boldsymbol{x}_i)$). The encodings are subsequently passed through a projection head, $Proj(\cdot)$, consisting of a single linear projection layer. The projection head generates lower-dimensional feature representation, denoted as $z_i = Proj(\boldsymbol{f}_i)$.

During training, the model, comprising the backbone encoder and the projection head, is optimized using supervised contrastive learning [21]. Specifically, for a given anchor sample $\boldsymbol{x}_i$, all augmented versions of $\boldsymbol{x}_i$ and samples from the same class as $\boldsymbol{x}_i$ are considered positive instances w.r.t. $\boldsymbol{x}_i$, while all other samples are treated as negative instances. The training objective is to encourage the model to draw the anchor sample closer to its positive instances in the latent space, while simultaneously pushing it farther from its negative instances.

The supervised contrastive loss function for the anchor sample is defined as:

$$\mathcal{L}_i^{cl} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(sim(z_i \cdot z_p)/\tau)}{\sum_{a \in A} \exp(sim(z_i \cdot z_a)/\tau)}, \quad (1)$$

where z_i, z_p, and z_a represent the latent representations of the anchor sample, a positive instance, and a negative instance, respectively. The function $sim(\cdot)$ computes the cosine similarity between two feature vectors, and τ is a temperature hyper-parameter that controls the sharpness of the distribution. The operator $|\cdot|$ denotes the cardinality of a set. The set of indices of positive samples with regard to the ith sample in a mini-batch is indicated by P_i, and the set of indices in the mini-batch that do not include i is indicated by A.

Training of the classifier. Following the first stage, the classifier is constructed by utilizing the trained encoder while discarding the projection head. A multi-layer perceptron classification layer is added on top of $Enc(\cdot)$. During this stage, the classifier and the encoder are fine-tuned simultaneously to classify input samples, employing the traditional CE loss as the objective function:

$$\mathcal{L}^{ce} = -\sum_{i}^{N} \log P(\boldsymbol{x}_i, y_i), \quad (2)$$

where $P(\boldsymbol{x}_i, y_i)$ denotes the probability of correctly classifying $\boldsymbol{x}_i$ as y_i.

Figure 3 compares the baseline model, which uses a binary classifier trained with CE loss, to the proposed

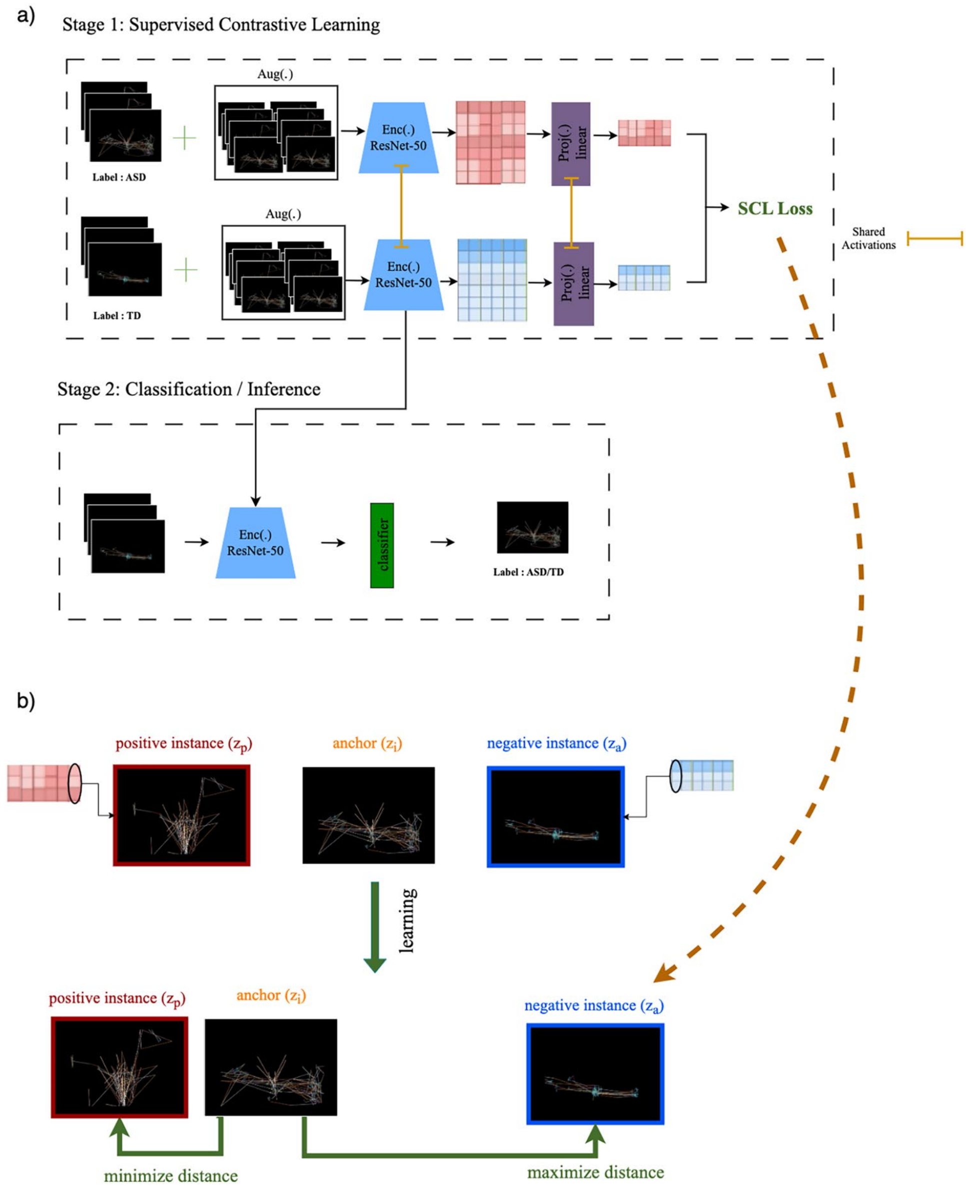


Fig. 2 **a** An overview of the proposed ConASD framework using ET scanpaths. $Aug(\cdot)$: Augmentation; $Enc(\cdot)$: Backbone encoder; $Proj(\cdot)$: Projection head. **b** Illustration shows how SCL is computed given an ET scanpath from ASD participant is treated as a positive instance z_p, while scanpath from different class (e.g., TD participant) is considered negative instance z_a for a given anchor sample z_i, within the ConASD framework

ConASD model. Both models operate under a FSL setting. ConASD uses SCL loss to optimize feature embeddings: the encoder output $\boldsymbol{f}_i = Enc(\boldsymbol{x}_i)$ is passed through the projection head $Proj(\cdot)$ to obtain z_i. Here, z_i' corresponds to the positive sample derived from an augmented version of the input, i.e., $x_i' = Aug(x_i)$. After training, the classifier predicts the label $\hat{y}_i$, which is compared with the ground truth y_i.

4 Experiments

This section presents our numerical experiments. We begin by describing the two datasets used in our study, followed by an outline of the baselines and experimental settings. Lastly, we present and analyze the results.

4.1 Datasets

The two datasets used in this study comprise ET based-images of children with ASD. They were selected to tackle a common limitation in medical research: datasets are often small due to the high cost of annotation, which poses challenges for effective model development.

Dataset 1. The first dataset was collected by [16]. ET experiments were conducted using an SMI Red-M eye tracker with a 60Hz sampling rate and a 17-inch monitor. Participants were seated 60-cm from the monitor in a quiet room with physical barriers to minimize distractions. Videos and other stimuli were employed to analyze eye activity from different perspectives. The experiments lasted approximately 5 min, during which participants' eye contact and focus were

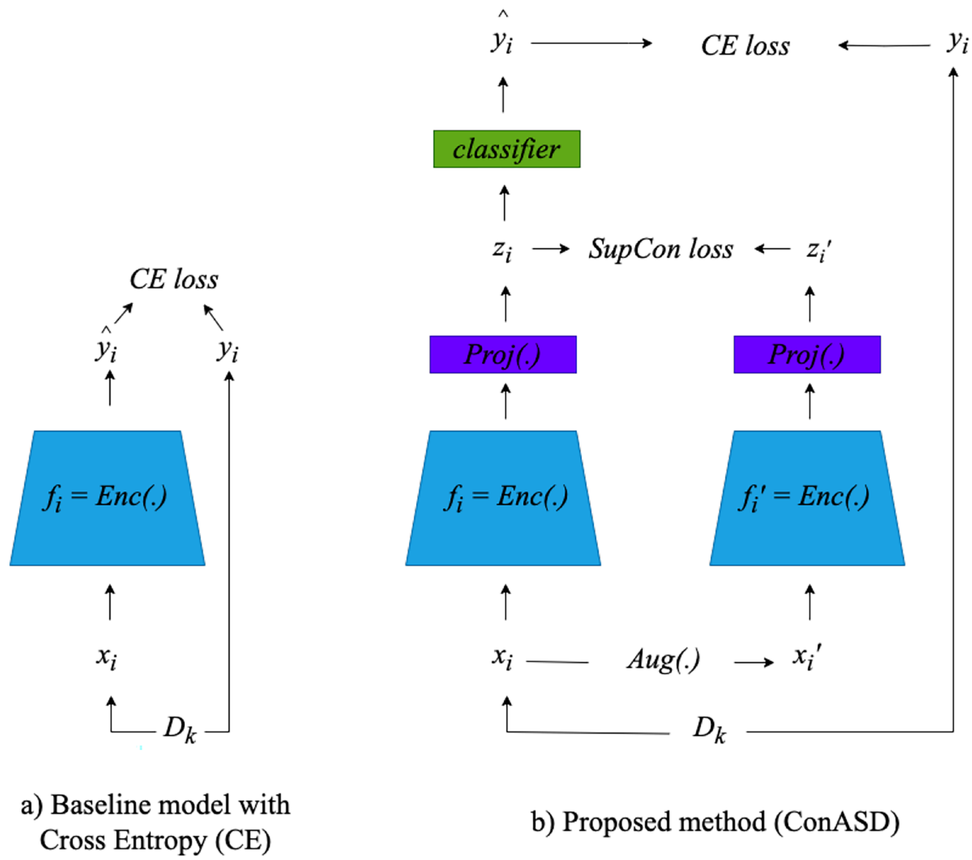


Fig. 3 **a** A comparison of the baseline method a using binary cross-entropy loss (Eq. 2), and **b** The proposed model (ConASD) that use supervised contrastive loss (Eq. 1). *Enc*(·): Backbone encoder; *Aug*(·): Augmentation; *Proj*(·): Projection head

closely monitored. A total of 25 ET experiments generated a dataset containing over 2 million records.

In ET, scanpaths are visual representations of gaze behavior, comprising sequences of fixations and saccades. Fixations denote brief moments of stationary gaze, while saccades represent rapid scanning movements. The collected ET scanpaths were converted into an image-based format by encoding gaze dynamics using color gradients. This process produced an image dataset of 547 images, including 328 images from TD participants and 219 individuals diagnosed with ASD. In Fig. 1, we illustrate four scanpath vizualisations: two from ASD participants and the other from TD children.

Dataset 2 The second dataset, consisting of eye movements of children with ASD-including scanpaths and fixation maps - was introduced in [40]. The data was collected from 14 ASD and 14 TD participants, capturing their gaze behavior in response to 300 natural scene images. Researchers used eye tracker called Tobii TX300, participants were seated 65-cm from the screen in a quiet, controlled environment to minimize distractions and ensure high-quality recordings. For the ASD group, all fixation points from participants diagnosed with ASD are combined into a binary map, where each fixation location is assigned a value of 1 and all other areas are set to 0. This process results in a group-level fixation map representing ASD visual attention. A similar method is applied to the fixation data from TD participants to produce their corresponding fixation map.

To improve visualization and interpretability, both binary fixation maps are smoothed using a Gaussian kernel, resulting in Fixation Density Maps, also referred to as visual attention maps. These maps capture the spatial distribution of gaze concentration across the visual field. The spread of the Gaussian kernel-reflecting the extent of visual attention-is typically defined by the standard deviation in degrees of visual angle. In Fig. 4, an example of a natural scene image viewed in the experiment is shown, along with two associated fixation maps: one for ASD children and the other representing TD children.

4.2 Baselines and settings

The evaluation is conducted for two ET datasets in the context of FSL with different training size scenarios. We evaluate the performance of our proposed method, ConASD, in comparison with four machine learning classification algorithms: Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and AdaBoost, as well as a CNN trained with CE loss. The ConASD and CE approaches use a ResNet-50 as the backbone

Fig. 4 Example of natural scene viewed to children (top), Bottom Left: fixation maps from TD children, Bottom Right: ASD children (sourced from [40])

for encoding input images. We select the ResNet-50 architecture over ResNet-101 based on the model results presented in Table 1. The results demonstrate that ResNet-50 delivers more stable and reliable accuracy with different training sizes and datasets.

The implementation is carried out using PyTorch, and experiments are conducted on a workstation equipped with an NVIDIA T4 GPU. Both datasets are split into training, testing, and validation sets following a 70%/20%/10% ratio. To simulate the FSL setting, we randomly sample training data, ensuring balanced datasets by selecting an equal number of samples (shots) from both the ASD and TD classes. Model optimization is performed using the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-4}. All models are trained for 100 epochs across all experimental runs.

4.3 Performance results

As outlined above, we trained the ET datasets on our proposed method ConASD based on contrastive learning and traditional methods which utilized CE loss under FSL settings. The comparative analysis of performance between two approaches shows notable observations, as presented in Table 2. In most experimental settings, with different training data and algorithms. The overall findings show that as the number of data size increases in few-shot examples, ConASD outperforms CE in terms of accuracy, precision, recall, and F1-score.

The results indicate that in scenarios with smaller FSL settings (e.g., 10 or 20 samples), CE occasionally achieves marginally higher accuracy. In some settings, SVM and LR show occasional improvements in precision and recall. However, ConASD demonstrates greater robustness and balance between precision and recall, resulting in higher F1-scores, which underlines its efficacy in handling limited training

Table 1 Performance comparison of ResNet-50 and ResNet-101 on two datasets with varying training sizes.

Training Size	ResNet-50		ResNet-101	
	Dataset 1 Acc	Dataset 2 Acc	Dataset 1 Acc	Dataset 2 Acc
10shot	0.6055	0.4500	**0.6514**	**0.5250**
20shot	**0.6514**	**0.5750**	0.6394	0.5417
50shot	**0.6697**	**0.6167**	0.4404	0.4833
100shot	**0.6881**	**0.6667**	0.5963	0.5667
200shot	**0.7328**	**0.7083**	0.5963	0.5666
Full Dataset	**0.6973**	**0.7667**	0.6147	0.6500

Bolded values indicate the highest accuracy achieved for each training size across the two architectures (ResNet-50, ResNet-101)

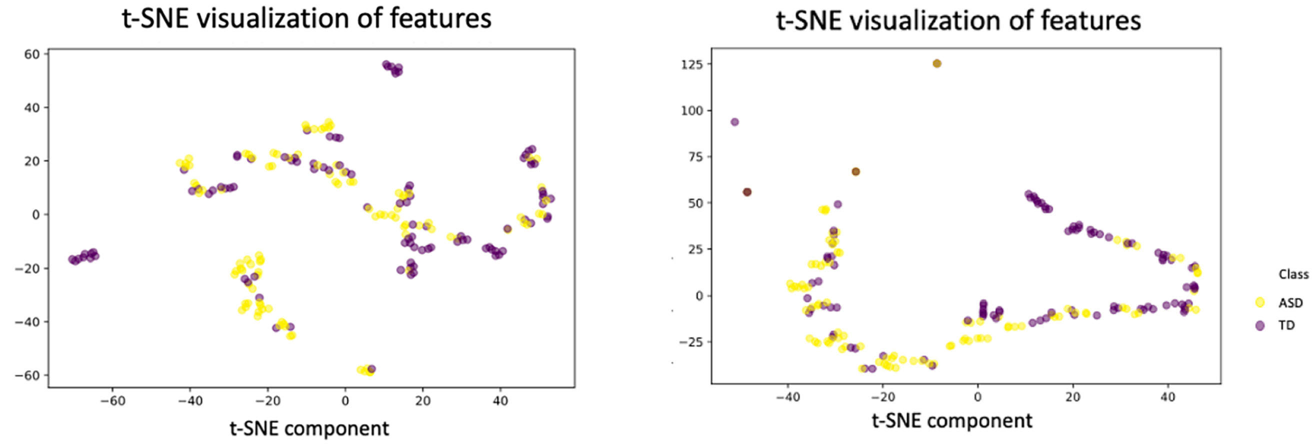
Table 2 Performance comparison of some ML baselines, ResNet-50+CE and ConASD in ASD detection across varying training data sizes – 10, 20, 50, 100, and 200, including results on the full dataset

Train size	Algorithm	Dataset 1				Dataset 2			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
10	DT	0.6239	0.6138	0.5561	0.5835	**0.6083**	**0.6086**	**0.6083**	**0.6085**
	SVM	0.6055	**0.8009**	0.5114	0.6242	0.5583	0.5622	0.5583	0.5603
	LR	0.6055	**0.8009**	0.5114	0.6242	0.5500	0.5521	0.5500	0.5510
	AdaBoost	0.5872	0.5329	0.5143	0.5234	0.6000	0.6004	0.6000	0.6002
	ResNet-50+CE	**0.6605**	0.7643	**0.5834**	**0.6617**	0.4750	0.3862	0.4750	0.4260
	Proposed ConASD	0.6055	0.6372	0.5638	0.6088	0.4500	0.4301	0.4500	0.4398
20	DT	0.6055	0.5815	0.5738	0.5776	0.5583	0.5604	0.5583	0.5593
	SVM	0.6147	**0.6798**	0.5264	0.5933	0.5083	**0.7521**	0.5083	0.6066
	LR	0.5963	0.5491	0.5037	0.5254	0.4917	0.2479	0.4917	0.3296
	AdaBoost	0.5780	0.5284	0.5177	0.5230	0.5667	0.5670	0.5667	0.5668
	ResNet-50+CE	**0.6606**	0.6334	0.5862	0.6088	**0.6750**	0.6867	**0.6750**	**0.6808**
	Proposed ConASD	0.6514	0.6734	**0.5921**	**0.6301**	0.5750	0.5760	0.5750	0.5755
50	DT	0.6239	0.6025	0.5928	0.5976	0.4500	0.4498	0.4500	0.4499
	SVM	**0.6697**	0.7030	0.6056	0.6507	0.5250	0.5252	0.5250	0.5251
	LR	0.6606	**0.7644**	0.5832	**0.6616**	0.5250	0.5252	0.5250	0.5251
	AdaBoost	**0.6697**	0.7030	0.6056	0.6507	0.5167	0.7542	0.5167	0.6132
	ResNet-50+CE	0.5872	0.6609	**0.6276**	0.6497	0.6083	**0.6478**	0.6083	**0.6274**
	Proposed ConASD	**0.6697**	0.6873	0.6199	0.6519	**0.6167**	0.6178	**0.6167**	0.6173
100	DT	0.5413	0.5168	0.5163	0.5165	0.5583	0.5585	0.5583	0.5584
	SVM	0.6606	0.6911	0.5942	0.6390	0.6083	0.6204	0.6083	0.6143
	LR	0.6697	0.7192	0.6019	0.6553	0.5417	0.5522	0.5416	0.5468
	AdaBoost	**0.6881**	0.7121	**0.6319**	0.6696	0.6167	**0.6389**	0.6167	0.6276
	ResNet-50+CE	0.6606	0.5604	0.5547	0.5575	0.6167	0.6178	0.6167	0.6173
	Proposed ConASD	**0.6881**	**0.7947**	0.6298	**0.7027**	**0.6667**	0.6794	**0.6667**	**0.6730**
200	DT	0.7064	0.6946	0.6878	0.6912	0.5666	0.5667	0.5666	0.5666
	SVM	0.7300	0.7247	0.7145	0.7196	0.6333	0.6347	0.6333	0.6340
	LR	0.7300	**0.7264**	0.7108	0.7185	0.6167	0.6234	0.6167	0.6200
	AdaBoost	0.7248	0.7150	0.7178	0.7164	0.7000	0.7002	0.7000	0.7001
	ResNet-50+CE	0.6789	0.6341	0.6233	0.6287	0.6333	0.6389	0.6333	0.6361
	Proposed ConASD	**0.7328**	0.7208	**0.7298**	**0.7253**	**0.7083**	**0.7089**	**0.7083**	**0.7086**
Full dataset	DT	0.6147	0.5895	0.5741	0.5817	0.6500	0.6515	0.6500	0.6508
	SVM	0.6514	0.6390	0.6049	0.6215	0.6583	0.6594	0.6583	0.6589
	LR	0.6514	**0.6508**	0.5939	0.6210	0.6500	0.6500	0.6500	0.6500
	AdaBoost	0.6514	0.6353	**0.6306**	**0.6329**	0.7083	0.7084	0.7083	0.7084
	ResNet-50+CE	0.6584	0.6348	0.6247	0.6297	0.6917	0.6961	0.6916	0.6938
	Proposed ConASD	**0.6973**	**0.6508**	0.5939	0.6210	**0.7667**	**0.7694**	**0.7667**	**0.7680**

The bolded results represent the best performance for each training size

dataset scenarios. Similarly, For the second dataset, ConASD demonstrates marginal improvements.

In addition, the results show that as the training dataset size increases, e.g., 100, 200 samples or the full dataset, the difference between ConASD and other baselines more clear. ConASD achieves higher performance metric confirming its ability to generalize to unseen data. In particular, with the largest dataset sizes, ConASD maintains consistently increased accuracy & F1 scores due to utilising contrastive learning, which improves feature representation in ET datasets. Thus, our findings demonstrate ConASD's effectiveness in FSL scenarios by maintaining a balanced trade-off between precision and recall, leading to improved F1 scores across various dataset configurations.



(a) CE (b) ConASD

Fig. 5 t-SNE vizualization of embeddings learned by CE and ConASD

4.4 Vizualization results

Figure 5 illustrates the t-SNE vizualization of embeddings learned by CE approach and our proposed method, ConASD, based on a dataset split of 70% training, 20% testing, and 10% validation, with the number of training samples set to 200 and the batch size set to 16. The yellow learned embeddings represent the ASD cluster, while embeddings in purple represent the TD cluster. A poor separation between the TD and ASD classes was indicated by scattered, overlapping clusters of the t-SNE vizualisation of CE approach. In contrast, ConASD produces compact, well-separated clusters, highlighting its ability to effectively capture distinct class representations. These findings are consistent with the quantitative results in Table 2, where ConASD outperforms CE in accuracy, F1-score and precision across most few-shot scenarios.

5 Conclusion

In this paper, we study the detection of autism spectrum disorder (ASD) through eye tracking (ET) images using deep learning models. In particular, we focus on the scenario where the training data is limited, which is common in ASD research where high-quality data is expensive to acquire. To address this challenge, we propose to use contrastive learning for extracting useful features from ET images for ASD diagnosis. We conduct experiments on two real-world datasets. The results confirm the advantages of ConASD in classifying ET's scanpaths for ASD and TD in FSL settings. Our framework, ConASD, consistently outperforms traditional ML baselines and CE approach for full dataset and most FSL scenarios. Such advantages are also confirmed by t-SNE vizualisations results, showing that ConASD generated more compacted, well-separated clusters, while CE produced scattered, overlapping clusters. This highlights the ability of contrastive learning in generating distinct and meaningful feature representations of visuals, which aligns with its superior quantitative performance. Overall, ConASD shows an effectiveness, especially in data-limited scenarios, addressing the challenges of similar cases in real-world medical data.

The proposed framework, ConASD, has few limitations. Firstly, contrastive learning relies on data augmentation, which may not be applicable to all types of ET data. Secondly, SCL requires a pre-training phase, making it more complex and costly compared to simpler methods like AdaBoost. Additionally, while the SCL is based on ResNet encodings, other architectures might be more effective. We plan to study these aspects in our future work.

Author contributions Estrid He contributed to the development of the presented idea. Sharifah Mousli wrote the first draft, conducted the experiments and prepared the figures. Sona Taheri and Estrid He participated in manuscript writing and review.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The data used in this research are publicly available and sourced from [16] & [40]. No Ethical approval required.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

1. Smith, T., Groen, A.D., Wynn, J.W.: Randomized trial of intensive early intervention for children with pervasive developmental disorder. Am. J. Ment. Retard. **105**(4), 269–285 (2000)
2. Jones, E.J., Gliga, T., Bedford, R., Charman, T., Johnson, M.H.: Developmental pathways to autism: a review of prospective studies of infants at risk. Neurosci. Biobehav. Rev. **39**, 1–33 (2014)
3. Majaranta, P., Bulling, A.: Eye tracking and eye-based human computer interaction. In: Adv. Physiol. Comput., pp. 39–65. Springer, Cham (2014)
4. Akter, T., Ali, M.H., Khan, M.I., Satu, M.S., Moni, M.A.: Machine learning model to predict autism investigating eye-tracking dataset. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 383–387 (2021). IEEE
5. Elbattah, M., Carette, R., Cilia, F., Guérin, J.-L., Dequen, G.: Applications of machine learning methods to assist the diagnosis of autism spectrum disorder. In: Neural Engineering Techniques for Autism Spectrum Disorder, pp. 99–119. Elsevier (2023)
6. Gaspar, A., Oliva, D., Hinojosa, S., Aranguren, I., Zaldivar, D.: An optimized kernel extreme learning machine for the classification of the autism spectrum disorder by using gaze tracking images. Appl. Soft Comput. **120**, 108654 (2022)
7. Hossain, M.D., Kabir, M.A., Anwar, A., Islam, M.Z.: Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. Health Inform. Sci. Syst. **9**, 1–13 (2021)
8. Kang, J., Han, X., Hu, J.-F., Feng, H., Li, X.: The study of the differences between low-functioning autistic children and typically developing children in the processing of the own-race and other-race faces by the machine learning approach. J. Clin. Neurosci. **81**, 54–60 (2020)
9. Mousli, S., Taheri, S., He, J.: Identifying autism spectrum disorder using optimization-based clustering. Open Sci. Index **17**(4), 44 (2023)
10. Stevens, E., Atchison, A., Stevens, L., Hong, E., Granpeesheh, D., Dixon, D., Linstead, E.: A cluster analysis of challenging behaviors in autism spectrum disorder. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 661–666 (2017). IEEE
11. Vargason, T., Frye, R.E., McGuinness, D.L., Hahn, J.: Clustering of co-occurring conditions in autism spectrum disorder during early childhood: A retrospective analysis of medical claims data. Autism Res. **12**(8), 1272–1285 (2019)
12. Cilia, F., Carette, R., Elbattah, M., Dequen, G., Guérin, J.-L., Bosche, J., Vandromme, L., Le Driant, B.: Computer-aided screening of autism spectrum disorder: eye-tracking study using data visualization and deep learning. JMIR Hum. Factors **8**(4), 27706 (2021)
13. Elbattah, M., Carette, R., Dequen, G., Guérin, J.-L., Cilia, F.: Learning clusters in autism spectrum disorder: Image-based clustering of eye-tracking scanpaths with deep autoencoder. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1417–1420 (2019). IEEE
14. Carette, R., Cilia, F., Dequen, G., Bosche, J., Guerin, J.-L., Vandromme, L.: Automatic autism spectrum disorder detection thanks to eye-tracking and neural network-based approach. In: Internet of Things (IoT) Technologies for HealthCare: 4th International Conference, HealthyIoT 2017, Angers, France, October 24-25, 2017, Proceedings 4, pp. 75–81 (2018). Springer
15. Carette, R., Elbattah, M., Cilia, F., Dequen, G., Guerin, J.-L., Bosche, J.: Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. In: HEALTHINF, pp. 103–112 (2019)
16. Cilia, F., Carette, R., Elbattah, M., Guérin, J.-L., Dequen, G.: Eye-tracking dataset to support the research on autism spectrum disorder (2022)
17. Ahmed, I.A., Senan, E.M., Rassem, T.H., Ali, M.A., Shatnawi, H.S.A., Alwazer, S.M., Alshahrani, M.: Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. Electronics **11**(4), 530 (2022)
18. Atyabi, A., Shic, F., Jiang, J., Foster, C.E., Barney, E., Kim, M., Li, B., Ventola, P., Chen, C.H.: Stratification of children with autism spectrum disorder through fusion of temporal information in eye-gaze scan-paths. ACM Trans. Knowl. Discov. Data **17**(2), 1–20 (2023)
19. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. arXiv preprint arXiv:1612.02295 (2016)
20. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR
21. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)
22. Yi, L., Liu, S., She, Q., McLeod, A.I., Wang, B.: On learning contrastive representations for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16682–16691 (2022)
23. Jones, W., Klin, A.: Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. Nature **504**(7480), 427–431 (2013)
24. Wan, G., Kong, X., Sun, B., Yu, S., Tu, Y., Park, J., Lang, C., Koh, M., Wei, Z., Feng, Z.: Applying eye tracking to identify autism spectrum disorder in children. J. Autism Dev. Disord. **49**, 209–215 (2019)
25. Elbattah, M.: Visualization of eye-tracking scan path in autism spectrum disorder: Image dataset. In: Proceedings of the 12th International Conference on Health Informatics, Prague, Czech Republic, pp. 22–24 (2019)
26. Bone, D., Goodwin, M.S., Black, M.P., Lee, C.-C., Audhkhasi, K., Narayanan, S.: Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J. Autism Dev. Disord. **45**, 1121–1136 (2015)
27. Ahuja, K., Bose, A., Jain, M., Dey, K., Joshi, A., Achary, K., Varkey, B., Harrison, C., Goel, M.: Gaze-based screening of autistic traits for adolescents and young adults using prosaic videos. In: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, pp. 324–324 (2020)
28. He, J., Li, Y., Zhai, Z., Fang, B., Thorne, C., Druckenbrodt, C., Akhondi, S., Verspoor, K.: Focused contrastive loss for

classification with pre-trained language models. IEEE Trans. Knowl. Data Eng. (2023)
29. Rethmeier, N., Augenstein, I.: A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. ACM Comput. Surv. **55**(10), 1–17 (2023)
30. Li, S., Xia, X., Ge, S., Liu, T.: Selective-supervised contrastive learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 316–325 (2022)
31. Rani, A., Yadav, P., Verma, Y.: Early-stage autism diagnosis using action videos and contrastive feature learning. Multimed. Syst. **29**(5), 2603–2614 (2023)
32. Auriau, P., Grigis, A., Dufumier, B., Louiset, R., Gori, P., Mangin, J.-F., Duchesnay, E.: Supervised diagnosis prediction from cortical sulci: Toward the discovery of eurodevelopmental biomarkers in mental disorders. In: 21st IEEE International Symposium on Biomedical Imaging (ISBI 2024) (2024)
33. Zhang, N., Ruan, M., Wang, S., Paul, L., Li, X.: Discriminative few shot learning of facial dynamics in interview videos for autism trait classification. IEEE Trans. Affect. Comput. **14**(2), 1110–1124 (2022)
34. Dhinagar, N.J., Santhalingam, V., Lawrence, K.E., Laltoo, E., Thompson, P.M.: Few-shot classification of autism spectrum disorder using site-agnostic meta-learning and brain mri. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1–6 (2023)
35. Wang, H., Chi, L., Zhao, Z.: Asdpred: an end-to-end autism screening framework using few-shot learning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 5004–5008 (2022)
36. Fawzy, A., Mossad, N., Ehab, Z., Moussa, N., Zaky, A.B.: Few-shot learning for early diagnosis of autism spectrum disorder in children. In: 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 89–92 (2024). IEEE
37. Zhou, H., Yang, L., Bao, H., Li, J., Li, Y., Miao, X.: Memristive cosine-similarity-based few-shot learning with lifelong memory adaptation. Adv. Intell. Syst. **5**(2), 2200173 (2023)
38. Bennett, C.C., Stanojevic, C., Oh, J., Ahn, J., Jeon, Y., Son, K., Abbott, N.: Digital health sensor data in autism: Developing few shot learning approaches for traditional machine learning classifiers. In: 2024 IEEE 20th International Conference on Body Sensor Networks (BSN), pp. 1–4 (2024). IEEE
39. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
40. Duan, H., Zhai, G., Min, X., Che, Z., Fang, Y., Yang, X., Gutiérrez, J., Callet, P.L.: A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the 10th ACM Multimedia Systems Conference, pp. 255–260 (2019)