Towards an automatic early screening system for autism spectrum disorder in toddlers based on eye-tracking

Daniel Fernandez-Lanvin[1] · **Martin Gonzalez-Rodriguez**[1] ·
Javier De-Andres[2] · **Raquel Camero**[3]

Abstract

According to official estimations, autism spectrum disorder (ASD) affects around 1% of European newborns. The high level of dependency of ASD-affected subjects entails an extremely high social and economic cost. However, early intervention can drastically improve children's development and thus reduce their dependency. One of the main common characteristics of subjects with ASD is difficulties with social interaction, which determines how they react to certain stimuli. This behavior can be automatically detected by analyzing their gaze. This study explores and evaluates the feasibility of automatic screening for ASD in toddlers under 24 months of age based on this specific behavior. We applied a matched pairs experimental design and a set of test videos, using a set of variables extracted from gaze analysis from toddlers using eye-tracking devices. The different videos try to capture social engagement, social information gathering gaze exchanges, and gaze following. We used the data to make a thorough comparison of machine learning algorithms (nine learning schemes), including some that were used in related prior research, and others that are popular in classification problems. The results show that several of the tested algorithms provided notable performance.

✉ Daniel Fernandez-Lanvin
  dflanvin@uniovi.es

  Martin Gonzalez-Rodriguez
  martin@uniovi.es

  Javier De-Andres
  jdandres@uniovi.es

  Raquel Camero
  raquelcamero.adansi@gmail.com

1 Department of Computer Science of the University of Oviedo, c/Leopoldo Calvo Sotelo 18, 33007 Oviedo, Spain

2 Department of Accounting of the University of Oviedo, Facultad de Economía y Empresa. Avenida Del Cristo s/n, 33006 Oviedo, Spain

3 ADANSI, 33213 Gijón, Spain

⚫ Springer

1 Introduction

The objective of the present study is to conduct a thorough evaluation of the performance of various machine learning algorithms for the prediction of autism spectrum disorder (ASD) in subjects under 24 months of age.

Epidemiological studies estimate that ASD affects around 1% of newborns in the European Union [1, 2]. Those affected by ASD develop a high degree of dependency that has significant family, social and economic cost. Buescher et al. [3] estimated that the costs of support services and interventions required by a person with ASD amount to more than a million euros. This cost is, however, related to the severity of the condition. Proper, early treatment of ASD has a strong influence on the development of the affected person, which is why the ability to diagnose this disorder as soon as possible is critical.

Such diagnoses currently require the intervention of an expert clinician. It is impossible to have these resources in all general pediatric clinics, and consequently the mean age of diagnosis is around 30 months old. Earlier diagnosis requires the development of objective tests that can be automated and do not need specialized staff. They could be performed by general pediatricians to identify and refer those at risk of ASD as early as possible.

One recurring strategy that previous studies [4–6] have shown to be effective in identifying ASD sufferers over 6 years old is the use of Eye Tracking devices to record subjects' reactions when interacting with certain social and non-social stimuli. These reactions in ASD toddlers differ from Typically Developing (TD) children, especially in the emotions and attention deficit they show [6, 7], two aspects than can be detected by analyzing subjects' gaze patterns. Eye Tracking allows an accurate record of an individual's gaze trace while observing a specific area. Some studies have produced promising results in the application of machine learning algorithms to ASD screening in older subjects [8–10]. However, all of those studies are based on samples from subjects over 24 months old. Furthermore, there are studies showing that gaze patterns shift with age for both ASD and TD subjects [11]. Finally, several of these studies are based on variables that cannot be extracted automatically.

The present study seeks to address that gap by investigating and assessing the feasibility of using gaze behavior to screen for autism spectrum disorder (ASD) in children under 24 months old. We found that several machine learning algorithms achieved high performance in diagnosis of ASD in newborns and toddlers based on their gaze behavior. As far as we are aware, this is the first study to give a reliable prediction in the diagnosis of autism in children under 24 months old, and this is the paper's key contribution.

The remainder of the paper is structured as follows: Section 2 outlines some general concepts about ASD and its diagnosis, while Section 3 discusses the most relevant prior research efforts. Section 4 explains the design of the study, including a description of the sample and the study variables, the filtering procedure, the machine learning algorithms to be tested, and the strategy for comparing their performance. Section 5 provides and discusses the results. Section 7.1 explains checks performed to guarantee the robustness of the results. Section 8 contains a summary and conclusions. Finally, Section 9 summarizes limitations and future research lines.

2 Early screening of ASD

According to the 5th Edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [12], ASD symptoms are conceptualized as an early-onset childhood developmental condition, when children are between 12 and 24 months old. Currently, these

symptoms are not detectable in earlier stages of childhood development. ASD can some-
times be detected at 18 months of age or even earlier. At 24 months old, a diagnosis made
by an experienced professional can be considered highly reliable [13]. However, the aver-
age age for definitive diagnosis is around 30 months, at which point ASD related behavior
is evident in most cases [12].

Fortunately, over the last 30 years, the incidence of associated issues related to ASD
(intellectual disability, lack of language skills or impossibility of educational inclusion) has
decreased significantly, mainly because of the implementation of early detection systems
[14]. Early intervention for people with ASD is a key tool when it comes to correcting
or compensating for the child's development. It has also proved to be of great value in
preventing deeper developmental changes and more severe disabilities [15]. It is accepted
worldwide that early detection and intervention in the disorder have clear positive effects
on the prognosis of ASD children [16].

Therefore, the great challenge is how to lower the screening age from the current
30 months to closer to the first appearance of symptoms (around 12 months). Early screen-
ing is not a trivial issue, since it is closely linked to detecting problems in the develop-
ment of linguistic and cognitive skills, which are difficult to assess in subjects younger than
24 months old. In reality, only about a third of parents of children younger than one year
old are aware that their children may exhibit ASD-related behaviors, but this percentage
climbs to between 80 and 90% for parents who are concerned about some aspects of their
child's development before they are 24 months old [17, 18].

Pediatric services monitor children's development between 12 and 24 months using
M-CHAT, an instrument to assess autism risk [19]. This is a questionnaire with 20 yes/
no questions that professionals complete based on their observations. Even though it is a
useful resource, it is not an objective test, which leads to a high number of false negatives
[20, 21]. In the event of a positive result, diagnostic tests are strongly recommended. The
M-CHAT is a screening tool, not a diagnostic tool.

On the other hand, experienced clinical professionals can make highly reliable diagno-
ses in children under 24 months old. They can identify risk indicators at around 12 months
old. These diagnoses are based on a qualitative assessments of children's visual, commu-
nicative, and motor development milestones or behavior that only a specialist with years
of experience working with ASD patients can recognize. This helps to explain why the
number of children diagnosed with ASD before the age of 18 months remains so low [15].

3 Prior studies

3.1 Eye-tracking analysis

As stated previously, one of the weaknesses of most effective screening tools for ASD
is that they rely on the observations of the development of cognitive and motor abilities.
Unfortunately, it is not possible to evaluate most of these skills in children who are under
24 months old. There are, however, some features that can be evaluated in this target group.
One of the most common behaviours in ASD subjects is an impairment in the ability to
make and maintain eye contact [22]. Eye tracking has been used to validate gaze-related
hypotheses about children with autism [23]. For instance, Jones and Klim [24] examined
eye fixation patterns in children from 2 to 6 months of age. Their results indicate that ASD-
diagnosed children exhibited mean decline in eye fixation that was not observed in those

who did not develop ASD. In addition, Chawarska et al. [25] noted that ASD children paid less attention to social scenes, people and faces. Wang et al. [26] used eye-tracking to demonstrate that people with ASD had a stronger image centre bias regardless of object distribution, and a reduced saliency for faces and for locations indicated by social gaze. The results of Pierce et al. [27, 28] showed that children with ASD aged 14 to 42 months preferred to visually examine dynamic geometric images rather than social images. Other studies have indicated that observation of the eye-gaze can be a key biomarker for diagnosing ASD [23, 29]. Vargas-Cuentas et al. [30, 31] developed both an algorithm and a low-cost tracking system based on a tablet and web-cam to track the gaze of children from 2 to 6 years old. Their results showed that children with autism spend less time watching social scenes. They worked with a sample of 31 subjects (23 TD control and 8 ASD subjects) and their results exhibited an error of 1.52% compared to the expert-based classification. One interesting feature of their proposal is that the subjects watched a set of five pre-recorded videos, making the evaluation more objective and potentially easier to automate than those based on interactions with parents or clinical staff. Carette et al. [22] developed a methodology to visualize the eye-tracking patterns of ASD diagnosed individuals, providing a dataset consisting of data from 59 participants between 3 and 13 years old, but without dealing with the automation of behavioral markers. Alie et al. [32] explored the possibility of using pattern recognition algorithms for infant gaze patterns at six months of age in children at high risk for ASD. They used video cameras and manual analysis of the recordings to track the gaze while children interacted with their parents, analyzing a binary variable (looking towards the face vs. looking away from the face) and 32 subjects. They used Variable-order Markov Models, obtaining an accuracy of 93.75%. Although their proposal is not a complete screening system, nor easily automated (it requires manual analysis of the videos and the parents' involvement), and was evaluated with a small sample (32 subjects, only 6 with ASD), their results are a promising indicator of the efficacy of this kind of test with young children. Tsang [33] analyzed the reaction of 55 Chinese children (25 TD control, 29 ASD) averaging 10 years old to several facial expressions using an eye-tracker device. Even though their results showed significant differences between ASD and TD subjects, the strategy is not applicable to younger children since it is based on expression recognition, a skill that is not present in children under 24 months old [34]. Murias et al. [35] conducted an experiment with 25 children from 24 to 72 months old to validate the strong association between gaze tracking of social communication outcomes with five well-validated caregiver-reported outcomes commonly used in clinical trials. Coco et al. [36, 37] performed an interesting study to evaluate and track the progression of ASD subjects. They designed a framework based on non-intrusive strategies analyzing ASD subjects' multiple visual cues using Facial Expression Recognition (FER) approaches, gaze tracking and head position. It was aimed more at improving ASD treatments than ASD screening. They tested the system with children aged 47–93 months (65.38 months average). This age range make sense in their study, since what they evaluate are mainly aspects related to subjects' social responses. However, these abilities are not developed in young toddlers. Vargas et al. [31] developed a low-cost table-based portable eye tracking system, with the aim of applying it to an ASD screening system. They created a set of 5 videos changing the social/abstract scenes and analyzed the gaze preference of the children for each social or abstract scene by placing them on the left or right side of the screen. They tested the system with a sample of 23 TD control children and 8 ASD children between 2 and 6 years old (48.71 months average age). Their results suggest they can distinguish with a high rate of success if the user is looking at the left or right side of the screen. However, this cannot be used for more fine-grained gaze analysis that may be necessary in this context, such as determining whether

a subject is looking at a speaker's eyes or mouth, or the number of saccades over each of these areas of interest (AOI). Because they have yet to create any automatic categorization method, the ASD screening is more of an incentive for the study than a true goal. A similar study was performed by Bovery et al. [7] in which they developed a low-cost alternative for measuring attention based on head and iris positions. This study is particularly interesting since they worked with 104 children (22 diagnosed with ASD) aged between 16 and 31 months old. In relation to attention to social versus non-social stimuli, they observed that the average number of frames looking at social versus non-social stimuli did not yield group differences. However, the more interesting conclusion for our study is related to the deficit in overall attention. They concluded that a participant paying attention to less than a certain percentage of frames would be one feature more commonly associated with ASD. With their results, considering 1,000 frames, the values of the precision, recall and F1 scores were P=0.8, R=1, and F1=0.89, respectively. They also found different patterns between ASD and control groups regarding attention shifting when watching the sample movies. These results are encouraging since, even though they are not determining criteria for ASD screening, they show that differences between ASD and TD toddlers can be detected automatically, even using low-cost devices. Nevertheless, Guillen et al. [6] noted that several studies show that this different behaviour between ASD and TD toddlers is strongly related to the context of the stimuli, which means that the results taken from the videos should be validated before being used as the basis for a classification system.

3.2 Machine learning (ML) based systems

The application of ML algorithms to medical screening is one of the most popular and effective current strategies [38], and ASD screening is no exception. The study by Hyde et al. [39] provides a comprehensive review of 45 papers using supervised machine learning in ASD, including algorithms for classification and text analysis. They concluded than the most widely used supervised machine learning algorithms were SVM and different Decision Tree algorithms. More recently, Jourdar et al. [40] described a systematic review of the literature to assess the respective AI methods using the available datasets, highlighting the tools and strategies used for diagnosing ASD. Most of those studies do not focus on eye-gaze tracking. Some, such as Lawi et al. [41], Tyagi et al. [42], Wingfield et al. [43], and Altay et al. [44] compare the accuracy of different ML algorithms over popular datasets based on questionnaires like M-CHAT (in these cases, the Autism Screening Adult Data Set by Thabtah[1]). Following a different approach, Sarabadani et al. [45] investigated detection of autonomic responses to positive and negative stimuli in 15 children with ASD using four physiological measurements (electrocardiograms, respiration, skin conductance and temperature). Krishnappababu et al. [46] compared the expressions of 40 ASD toddlers with 396 TD toddlers watching certain videos. They analyzed subjects' eyebrows and mouth regions, finding significant differences in the response of ASD subjects.

The classification approach in these studies is essentially different than the approach in our study. Nevertheless, we believe that their experience with the different ML algorithms can be useful in driving our research.

That said, some previous studies are far more interesting for our specific context, since they apply ML to the specific processing of gaze tracking data. Wan et al. [47] evaluated

---

[1] http://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

the fixation times of 37 ASD and 37 TD children aged 4 to 6 years watching a 10-s video of a woman speaking. Their results, based on an SVM classification system, suggest that "a short video clip may provide enough information to distinguish ASD from TD children" [47]. These results are encouraging but are based on a sample of subjects that are too old to be able to extrapolate the results to the target age for our study. Dris et al. [15] developed what they consider a "proof of concept" combining gaze-based screening with SVMs to diagnose ASD subjects. They reported an accuracy of 86%. However, there is no information about sample size or age distribution, nor any definitive design or conclusions. Canavan et al. [9] used gaze information from the National Database for Autism Research[2] to create three different classifiers based on C4.5, Random Forest and Partial Decision Tree (PART) algorithms, concluding that the latter was the most effective with 96.2% accuracy. The sample included subjects aged from 1 to over 60 years old. Jiang et al. [8] examined the feasibility of using eye movements combined with task performance in facial emotion recognition to identify individuals with ASD. They applied Random Forest to a sample of 23 subjects with ASD and 35 TD controls with an average age of 12.74 years, obtaining a classification accuracy of 86%. Although these are promising results, the fact that the experiment was based on face recognition makes it impossible to apply it to children under 24 months old. Previously, the same team had conducted another experiment focused on ASD screening using eye tracking and deep neural networks [4] with 39 subjects, using an image dataset and heat maps for each sample member. Their results (92% accuracy) indicate the effectiveness of the classification algorithm for the specific problem.

A very interesting and novel approach is that proposed by Oliveira et al. [48]. Their work addressed the ASD screening challenge with a machine learning approach to dispense with the use of AOIs and develop a classifier based on Visual Attention Models (VAMs) learned for each group of individuals. The main difference is that instead of determining the discriminating AOIs according to the clinical experts, they create a group-specific fixation map for the videos identifying the VAMs (VAM learning phase) using ML algorithms. That makes the system more flexible, since new videos can be used for training and classifying, but also potentially more accurate, since the use of manually determined AOIs may not fit exactly with the actual behaviour of both ASD and typically developing subjects. Their results are already promising (90% precision, 69% recall and 93% specificity), but their sample of 106 subjects is made up of 3 to 18 year-olds, outside the target age range for our study.

Akter et al. [49] evaluated different machine learning algorithms to try to predict ASD in subjects aged 3 to 13 years old. They analysed the scan path images collected from the Figshare data repository, and classified them using k-means clustering, generating 4 clusters of subjects. They then evaluated several ML algorithms (Decision Trees, Gradient Boost, KNN, Logistic Regression, Multilayer Perceptron, Naïve Bayes, Random Forest, SVM and XGBoost). Their results indicate that the best performance was with the Multilayer perceptron with the first of the clusters (87% accuracy). Although the procedure and results are interesting, they do not explain the kind of stimulus that ASD and TD subjects preferred, nor the differences in the pictures that underlay the appearance of the 4 clusters. These results reinforce the accepted idea that ASD subjects have a different gaze pattern when faced with certain stimuli, and that ML algorithms can detect these differences automatically. However, they are evaluating classification algorithms in clusters that, at the

---

[2] https://nda.nih.gov/about.html

same time, have been already created with unsupervised ML algorithms, which makes it difficult to apply to generalised ASD screening. In addition, their sample age range is outside our target range.

Despite the intensive activity in this research topic, none of the studies analysed proposed an effective, tested, automated, objective method to support the early screening of ASD in young children. Nonetheless, these studies do provide useful information about the different ML algorithms, indicating their effectiveness for this specific problem. Table 1 summarizes the algorithms tested by these studies, those specifically focusing on gaze analysis are highlighted in grey.

4 Design of the empirical study

4.1 Methodology

Figure 1 depicts the methods used in this study. Each participant was exposed to a series of visual stimuli while their gaze was tracked (step 1).

The raw data obtained by the eye tracker was then automatically analysed and filtered (step 2) using algorithms to remove noise (the consequence of erroneous eye tracker recordings). The data were processed to obtain values of high semantic level. These variables include, for example, the number of gaze exchanges between areas of interest (AOI) linked with people (mouth, eyes, fingers) and others connected with objects (puppets, toys, etc.), which will be detailed later.

Step 3 involved screening every volunteer for ASD using clinical processes, resulting in two potential categories: ASD and non-ASD. This process was done manually by specialists. It is important to note that reliable ASD diagnoses cannot be obtained until a child is well over 3 years of age (and even older), at which point the behaviours linked with the disorder are manifest. As this was a pioneering study focusing on children under the age of 24 months, it took months in some cases and years in others to get a robust categorization for each participant.

Step 4 involved analysing the collected values using a correlation feature selection (CFS) approach to identify variables of high importance that might contribute to the development of effective classification systems for this context. This procedure is described in subsection 4.5.

In step 5, the successfully classified samples (ASD and TD) were used to train a series of classification models based on 9 Machine Learning algorithms (explained in subsection 4.6). Results were then analysed using various metrics to validate their efficiency and applicability to this problem domain.

4.2 Infrastructure

To the best of our knowledge, this study is the first of its kind, focusing on identifying and evaluating ASD indicators in a very young age group. Given the unique developmental characteristics of babies and toddlers, the creation of a specialized data acquisition environment was deemed necessary to facilitate accurate and efficient screening processes.

A customised chair was created for this purpose to keep newborns and toddlers comfortable during the observation session and always at the exact distance required by the eye tracker. To calibrate the eye tracker (a necessary step at the start of each session with
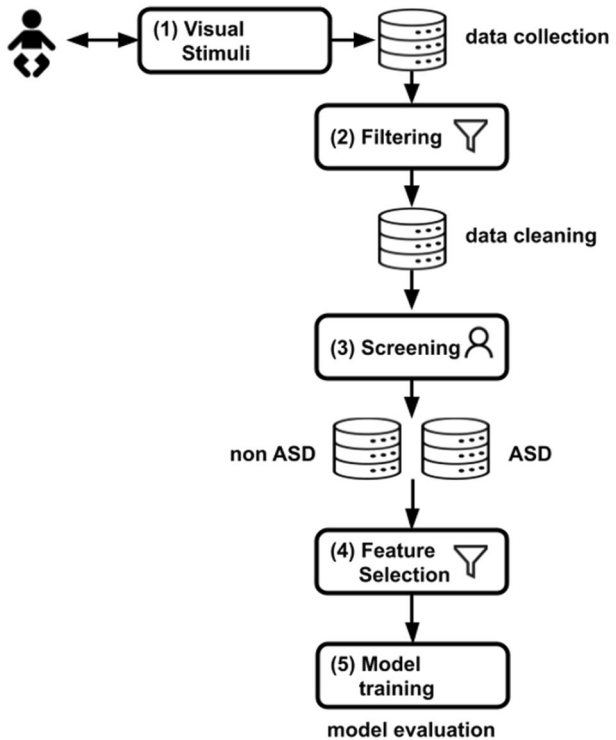
Table 1 ML algorithms tested in ASD screening and evaluation. Percentages indicate accuracy

| | Decision trees | | | | Linear statistical models | | | | | | |
	Random Forest	C4.5 Quinlan	PART	Classification And Regression Trees (CART)	Linear Discriminant Analysis	Linear Regression (LR)	Naïve Bayes (NB)	K-Nearest-Neighbor (KNN)	Support Vector Machine (SVM)	Variable-order Markov Models (VMM)	Neuronal Network
Non Eye-gaze tracking data based works											
Jiang et al. [4]	96%										
Lawi et al. [41]		100%			96.99%	99.6%		92.99%	99.9%		
Drimallla et al. [10]	0.84 (AUC)								0.81 (AUC)		0.64 (AUC)
Tyagi et al. [42]				69.17%	72.2%	72.02%	70.78%	67.56%	70.59%		
Altay et al. [44]					90.8%			88.5%			
Alie et al. [32]										93.75%	

Table 1 (continued)

		Decision trees				Linear statistical models							
		Random Forest	C4.5 Quinlan	PART	Classification And Regression Trees (CART)	Linear Discriminant Analysis	Linear Regression (LR)	Naïve Bayes (NB)	K-Nearest-Neighbor (KNN)	Support Vector Machine (SVM)	Variable-order Markov Models (VMM)	Neuronal Network
Eye-gaze tracking data based works	Jiang et al. [4]											92%
	Jiang et al. [8]	86%										
	Dris et al. [15]									86%		
	Wan et al. [47]									85.1%		
	Canavan et al. [9]	93.25%	94.94%	96.2%								
	Oliveira et al. [48]									0.775 (AUC)		0.822 (AUC)
	Akter et al. [49]											88,89%

Fig. 1 Flowchart for test design

the device), the speech therapist carefully selected specific cartoons to hold the children's attention. These graphic components were presented in different parts of a large screen that was specifically constructed for this purpose, allowing the children to comfortably watch them without turning their heads during the calibration phase before data collection.

A high-performance eye tracker capable of capturing gaze data at speeds up to 1200 Hz was selected in anticipation of normal newborns' quick eye movements. The Tobii Pro Spectrum model was selected, with Tobii Lab Pro software used to handle raw data gathering.

This new infrastructure was built at the ADANSI facilities, an association of relatives of people with autism located in the Principality of Asturias (Spain).

4.3 Stimuli and variables

The stimulus battery was meticulously prepared with such young participants' perceptual needs in mind. To accomplish this, a script for a series of very short films was created, which were shot with a combination of actors, puppets, and computer-generated images (CGI) added in post-production. To minimise the habituation effect and the resulting possible bias, the visual protagonists (both human and inanimate) switch locations in each film (see Figs. 2, 3, and 4). The movies were created to elicit typical visual reactions in the gaze of a typically developing non-ASD, child to the many protagonists, as well as to their components (eyes, mouth, hands, and so on). The AOIs related to both social and non-social stimuli present in each scene in the videos were determined to anticipate this behaviour.
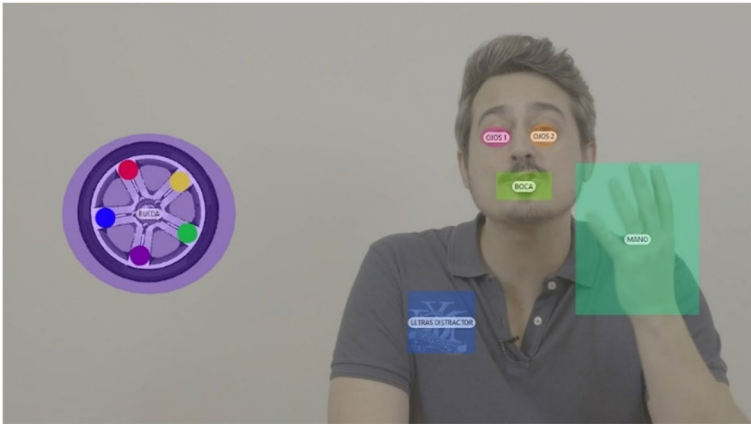
Fig. 2 Example of human figure and distracting element (a wheel)

Fig. 3 Example of "social information gathering gaze exchanges" video

The movies were created as part of ADANSI's "Cómo mira tu bebé" ("How your baby looks") Project for early diagnosis of autism risk, which was based on the design by psychologist Gloria Acevedo Díaz [49].

The short movies used as the stimulus can be divided into 3 categories based on the variables being analysed.

The first category is "social engagement" and consists of three videos. The attentional response is measured to a social stimulus and to a distracting element (which is not of a social nature, for example, an object). This shows the child's inhibitory capacity and attentional preference towards social stimuli.

To create this category of videos and specifically apply them to ASD screening, we reviewed previous studies conducted with eye-tracking systems where results of altered gaze patterns were obtained for children at risk of being diagnosed with ASD. Shic et al.

Fig. 4 Example of a "gaze and deictic tracking" video
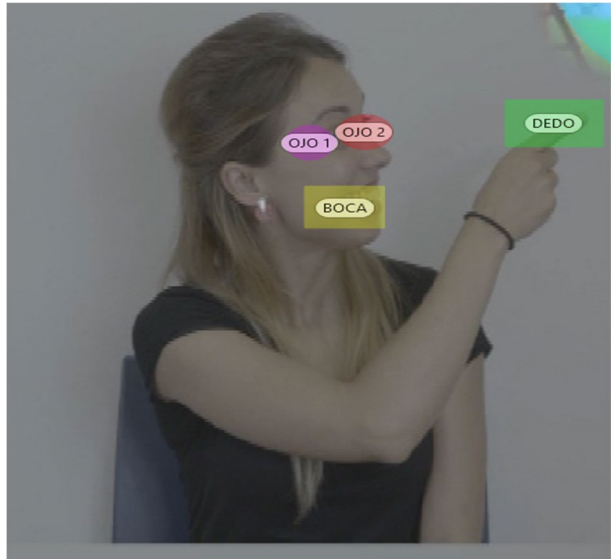
[50] found abnormal attentional patterns for human face processing in six-month-old babies who were later diagnosed with ASD. As noted in Section 3, Pierce et al. [27] stated that two-year old children diagnosed with ASD showed more preference in fixing their attention on geometric shapes rather than on human faces.

In each video, a human figure appears and seeks the child's attention by singing, clapping, or repeating phrases with intonation typical of child-directed speech. It should be noted that in the first video there are only two points at which the human face and the distracting element appear in different camera planes to measure the child's attentional behaviour with the objectives separately. In the rest of the videos, the stimuli appear in the same camera plane.

The variables measured are the child's reaction time (time it takes to fix attention on the human face), the time and number of fixations towards the eyes and mouth of the social element (human face) and towards the distracting element (object).

Figure 2 shows an example image from the video. The actor interacts with the subject by singing a typical kindergarten song. The AOIs are highlighted in colour. On the right, the human figure with the areas: eyes, mouth, hand, and a distracting element of the shirt such as the letters, and on the other side, the main distracting element, which is the moving wheel to the left of the actor. (Image courtesy of ADANSI).

The second category is "social information gathering gaze exchanges" and comprises three videos. A human figure and an object appear again but the human figure refers to the object (balls, numbers, and animal toys) during the video. It reproduces the real situation of adult–child interaction and is how the child's divided attention capacity between the social stimulus and the object can be analysed.

This group of videos is based on the difficulty children with ASD have in processing social information coming from a context. These children also have difficulty switching their attention between an object and the person that is talking about the object [51]. The variables measured are time and number of fixations to the eyes and mouth of the human face and to the different toys that appear. In addition, the number of exchanges between the

different AOIs on the human face (eyes or mouth) and towards the object are also measured. We count an exchange each time the subject's gaze moves from any face AOI to the object and vice versa (completing the bidirectional trajectory).

Figure 3 shows an example of a human figure-object interaction (balls). The actress plays a game with a ball ramp toy, moving her gaze between the different balls and looking directly at the camera, so that the observer feels that the actress is talking directly to them. The areas of interest are marked in colour (the human figure (eyes and mouth) and the object (balls) (Image courtesy of ADANSI)).

The third category is "gaze and deictic tracking" and consists of four videos. A human figure appears pointing or looking in a certain direction where some cartoons appear. The first video of this category is used as training, to familiarise the child with the task. The human figure points to different areas in the camera plane where the different stimuli (targets) appear. The second video is similar, but with an added waiting time. This allows us to measure whether the child can follow the direction of the human figure's finger before they see the target. In the third and fourth videos, the human figure does not point, she just turns her head and looks in a specific direction (target).

This set of videos is based on the paradigm that children with ASD are not able to follow what an adult points at, or follow the direction of an adult's gaze. This is a prerequisite for developing "joint attention" and is basic to language learning and social skills, which are critical for children's cognitive development [52].

The variables measured are the child's reaction time (time taken to fix attention on the human face); the time and the number of fixations towards the person's eyes, mouth, and pointing finger, and towards the different cartoons that appear. In addition, the number of visual trajectories that the child exchanges towards the human face (eyes or mouth) and towards the cartoons were also measured. We also defined a binary variable (visual fixation/absence of visual fixation), measuring the child's ability to look towards the place indicated or observed by the adult.

Figure 4 is an example image from a gaze following video. The actress points to and looks at different parts of the screen where a series of drawings appear. There is a waiting time when she points and looks in a certain direction, but nothing appears immediately. If the child looks in the right direction before the cartoon appears, it is considered a hit. The AOIs appear in colour (the human figure (eyes, mouth, finger), the empty areas during waiting times and the different drawings that appear that would be considered attention to the object) (Image courtesy of ADANSI).

The total set of initial variables comprises 132 indicators. The initial set of 132 variables was acquired directly by the Tobii Pro Lab software during each observation session with the Eye Tracker. The values were obtained in comma-separated value text files. The following procedure was used to assign variable names:

$$\mathrm{vN}\_\dfrac{\dfrac{tr}{tf}}{\dfrac{nf}{ni}} + \dfrac{social\ item}{object\ item}$$

vN_(tr/tf/nf/ni)(social item/object item/other)

V stands for "video" and N refers to the number of the video. Then, the quantitative type of measure is added: rt means the child's reaction time, tf means time of fixation to a certain element, nf is the number of fixations to a certain element and ne is the number visual

trajectories exchanged between two elements. Finally, the name of the specific element is added (defined areas of interest). It may be a social element (mouth, eyes, finger) or an object (cartoon, earrings, balloons, doll, number, wheel, dog, generic object) or binary variables defined in the third category (error and empty).

For example, v1_nfmouth. This variable is extracted from the first video (1) and measures number of fixations (nf) to the mouth. Another example, v5_neeyemouth: this variable is extracted from the fifth video (5) and measures number of visual trajectories exchanges (ne) between two social elements (eye and mouth).

We also defined a binary variable, ASD, which is one if the subject has ASD and zero otherwise. This variable is the dependent variable in the different classification devices.

4.4 Data collection

The data collection project was approved by the Principality of Asturias IRB (Institutional Review Board), the Research Ethics Committee of the Principality of Asturias (Protocol number: 2021.260).

All the children who took part in the experiment watched the videos, and detailed records of their gaze trajectories were retained for later analysis. The data was then processed using filtering algorithms (see step 2 in Fig. 1) to obtain gaze trajectories, fixations (periods in which the eyes are locked towards a specific AOI), fixation sequences, time to first fixation, time between fixations, time spent on each AOI, saccades (eye movements between fixations), and so on.

The study used a matched pairs design, so each of the ASD subjects was paired with a TD child of the same age. Subjects were assigned to the ASD group following the ADANSI experts' criteria: not only by clinical diagnostic criteria, but also by their scores above the cut-off line in the ADOS-2 Observation Scale for Diagnosis of Autism [53]. The matching criterion was the developmental age on the Brunet-Lezine Early Childhood Psychomotor Development Scale [54].

The final sample comprised 122 children. The breakdown of the sample by age and the main descriptive statistics are shown in Fig. 5.

The datasets produced and analysed during the study are not publicly available due to the nature of the information, but are available from the corresponding author on reasonable request.

4.5 Variable filtering

Considering that there were more variables than individuals in the sample, we applied a filtering process to discard non-relevant variables. To do this, we used a correlation feature selection (CFS) process based on entropy measures to retain only the variables that were correlated to a certain extent with the class, but not with other features in the selected subset. Ultimately, the selection of a feature is determined by the extent to which it predicts classes in areas of the instance space not predicted by other features. CFS ranks feature subsets according to a heuristic evaluation equation, which takes the following form:

$$M_S = \frac{k \times \overline{r_{cf}}}{\sqrt{k + (k(k-1) \times \overline{r_{ff}}}}$$

Fig. 5 Descriptive statistics and breakdown of the sample by age
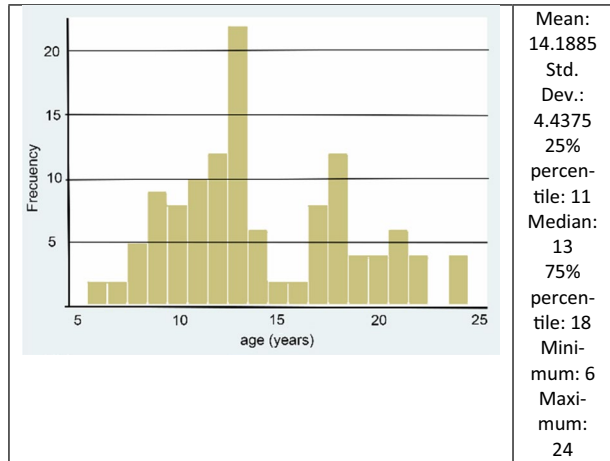
where M_S is the heuristic "merit" of a feature subset S containing k features, estimated r_{cf} is the mean feature-class correlation (f∈S), and estimated r_{ff} is the average feature-feature intercorrelation. A full description of the process, including the way correlation is computed using an entropy measure which is explained in Hall [55] and Doshi and Chaturvedi [56].

We applied this procedure using a stratified tenfold cross-validation strategy. Under this approach, the data were randomly split into 10 mutually exclusive subsets of approximately equal size. Each of the subsets replicated the class proportions existing in the global sample. Feature-class correlations and feature-feature-intercorrelations were computed having into account fold partitions, and the heuristic M_S score described above was computed for each of the feasible feature subsets. Then, the best subset (the one with the highest M_S score) was identified. We retained only the variables that were in the best subset in at least one of the folds.

The final list of variables, made up of 37 indicators, is shown in Table 2, which also includes a brief description of each one. Appendix 1 expands on this information, giving the mean, standard deviation, and maximum value (minimum is zero in all cases) for each variable by group (ASD and TD), as well as the significance of a paired t-test for the difference of means.

The results of the t-tests showed that for some of the variables the difference of means was either not significant or significant only at the 10% level. We decided to retain those variables as they may be useful for classification in a multivariate context.

4.6 Machine learning algorithms

We used several machine learning algorithms to construct a series of classifier systems. The chosen algorithms are either standard in the field of statistical learning or have been used in previous related research:

Table 2 Variables used to construct the classifiers

Variable	Description
vid1_react_time	Reaction time when shown video 1
vid1_num_fix_mouth	Number of fixations on the mouth of the person in video 1
vid1_num_fix_balloons	Number of fixations on the balloons in video 1
vid1_num_fix_eyes	Number of fixations on the eyes of the person in video 1
vid1_time_fix_balloons	Time of fixation on the balloons in video 1
vid1_time_fix_eyes	Time of fixation on the eyes of the person in video 1
vid2_num_fix_eyes	Number of fixations on the eyes of the person in video 2
vid2_num_fix_doll	Number of fixations on the doll in video 2
vid2_time_fix_eyes	Time of fixation on the eyes of the person in video 2
vid2_time_fix_doll	Time of fixation on the doll in video 2
vid3_num_fix_wheel	Number of fixations on the wheel in video 3
vid3_time_fix_mouth	Time of fixation on the mouth of the person in video 3
vid4_num_fix_mouth	Number of fixations on the mouth of the person in video 4
vid4_num_fix_earrings	Number of fixations on the earrings of the person in video 4
vid4_num_fix_number	Number of fixations in the number appearing in video 4
vid4_time_fix_mouth	Time of fixation on the mouth of the person in video 4
vid4_time_fix_earrings	Time of fixation on the earrings of the person in video 4
vid4_time_fix_number	Time of fixation on the number in video 4
vid4_num_ex_eyesmouth	Number of exchanges between the eyes and the mouth of the person in video 4
vid5_num_ex_eeyesmouth	Number of exchanges between the eyes and the mouth of the person in video 5
vid6_num_fix_eyes	Time of fixation on the eyes of the person in video 6
vid6_time_fix_mouth	Time of fixation on the mouth of the person in video 6
vid6_time_fix_eyes	Time of fixation on the eyes of the person in video 6
vid6_num_ex_eyesmouth	Number of exchanges between the eyes and the mouth of the person in video 6
vid6_num_ex_eeyesgenericobject	Number of exchanges between the eyes of the person and the generic object in video 6
vid7_react_time	Reaction time when shown video 7
vid7_num_ex_eyesmouth	Number of exchanges between the eyes and the mouth of the person in video 7
vid7_num_ex_eyescartoon	Number of exchanges between the eyes of the person and the cartoon in video 7
vid7_num_ex_fingercartoon	Number of exchanges between the finger of the person and the cartoon in video 7
vid8_num_fix_empty3	Number of fixations on the 3rd empty space in video 8
vid8_time_fix_error	Time of erroneous fixation in video 8
vid8_num_ex_eyesfinger	Number of exchanges between the eyes and the finger of the person in video 8
vid8_num_ex_eyesmouth	Number of exchanges between the eyes and the mouth of the person in video 8
vid8_num_ex_mouthcartoon	Number of exchanges between the mouth of the person and the cartoon in video 8
vid9_time_fix_empty2	Time of fixation on the 2nd empty space in video 9
vid10_time_fix_dog	Time of fixation on the dog in video 10

Table 2 (continued)

Variable	Description
vid10_num_ex_mouthdog	Number of exchanges between the mouth of the person and the dog in video 10

4.6.1 C4.5 and PART

The C4.5 and PART algorithms were used to build the decision tree, which is a predictive machine learning model. Quinlan [57] has advanced the ID3[58] technique with his C4.5 model for the induction of decision trees, that has been applied to a variety of classification problems [9, 41]. In this research we used the J48 algorithm, a Java implementation of the C4.5 algorithm.

PART [59] is an evolution of the C4.5 algorithm which uses the "separate and conquer" strategy to build decision lists. It has been used in several prior studies on ASD prediction [9].

4.6.2 AdaBoost

AdaBoost (Adaptive Boosting) [60] is a meta-learner, a learning algorithm that is applied to the results of machine learning experiments. It starts from a collection of "weak" learners, which are combined to form a "strong" classifier. In this study, as weak learners, we used decision stumps— which are one-level decision trees. We used an entropy measure to estimate the decision stumps.

Furthermore, this is a boosting algorithm, which means that a set of weights over the original training set is maintained, and these weights are adjusted after each classifier is learned by the base learning algorithm. In this study, we used boosting by weighting, meaning that the entire training set and associated weights are given to the base learning algorithm. As several authors have indicated (e.g., [61]) AdaBoost is suitable for problems in which training and test data lack noise (i.e. examples with incorrect class labels), which is the case in this study.

4.6.3 Random Forests

Random Forests [62, 63] is a method that, when applied to classification tasks, grows a number of trees using a base algorithm and does prediction by using a voting scheme.

The base algorithm in this study was the Reduced Error Pruning Tree (REPTree), which is a modification of C4.5. We set the number of trees at 100. In addition, in the Random Forests algorithm the base learner is modified so when growing the tree only a subset of the features randomly selected is considered at each node,. A common procedure, which we followed, is to set the number of randomly chosen attributes at $int(\log_2(\text{number of features}) + 1)$.

The Random Forests method has been extensively used for classification tasks in a wide variety of domains, and specifically in ASD screening as shown in Table 1 [4, 8–10]. Its advantages include the fact that it can avoid overfitting in most of cases and is relatively robust against outliers [64].

4.6.4 Support Vector Machines (SVM)

SVMs [65] provide a linear model which attempts to learn the maximum margin hyperplane that separates two categories in the supplied data.

As some authors have indicated (e.g. [66]), SVM offers good results when specialized prior knowledge about a certain domain is scarce. This is the case of the present problem, as we do not a priori know which of the considered variables are more important for the ASD classification problem, which several authors have applied it to successfully [10, 15, 41, 42, 47]. Furthermore, as in most cases the number of support vectors is far lower than the number of examples, SVMs gain some of the advantages of simple linear parametric models.

4.6.5 Naïve Bayes

This algorithm belongs to the family of Bayesian Networks and is based on application of Bayes' theorem with the assumption of strong independence between the features used for classification [67]. We used a kernel estimator to estimate probability density functions, as it has proven to provide higher accuracy levels [68].

Despite its simplified assumptions, Naïve Bayes has yielded good results in several classification problems, including some related to medical screening, and specifically, ASD [42]. There are a number of reasons for the efficacy of the Naïve Bayes classifiers, including that they only need a small amount of data to estimate the parameters necessary for classification [69]. This is crucial in our research problem, as obtaining new data is costly.

4.6.6 K-Nearest Neighbour (k-NN)

This method finds the instances that are nearest to the one to classify, and then takes the plurality vote of the neighbours [70]. To apply this method, two issues must be adequately addressed. First, a distance metric must be chosen. We used the Euclidean distance. Second, the number of neighbour instances must be determined. We used a hold-one-out cross-validation procedure to select the best k value. Like other methods, k-NN has been previously used for ASD-related studies [41, 42, 44].

4.6.7 Neural Networks

Like several prior studies on ASD prediction [4, 10], we included a Neural Network model in the set of classifiers. Specifically, we used the backpropagation algorithm to learn a multi-layer perceptron. The nodes in our network model are all sigmoid. Neural Networks have proven to provide good nonlinear separation capabilities. They have been extensively used in topics where there is no exact knowledge about the functional form of the relationship between variables.

4.6.8 Logit

Finally, to offer proper benchmarking, we computed the results of a linear model. Although some previous studies have also used linear discriminant analysis, we decided to use only logistic regression due to its less restrictive assumptions.

For the calibration of the different ML methods, we used the meta-classifier opti-mizer based on Bayesian optimization proposed by Kothoff et al. [71] and implemented in the WEKA environment as an external package.

4.7 Comparison strategy

To compare the performance of the different ML algorithms we used a stratified tenfold cross-validation strategy which we repeated 200 times, so for each of the calculated per-formance indicators we had 2000 observations. A first group of indicators was made up of those produced from the confusion matrix: sensitivity, recall, hit rate, or true positive rate (TPR); specificity, selectivity, or true negative rate (TNR); miss rate or false nega-tive rate (FNR); fall-out or false positive rate (FPR); and the total percentage of correct classifications (TPCC).

However, we must also bear in mind that the ASD classification problem has asym-metric misclassification costs, but the cost distribution is unknown. So to conduct statis-tical tests on the performance of the different algorithms, we used the receiver operating characteristic (ROC) curve, which plots TPR on the vertical axis against TNR on the horizontal. The area under the curve (AUC) can be understood as the probability that the classifier ranks a randomly chosen positive individual above a randomly chosen neg-ative one, so it can be used to assess which classification scheme performs better when considering all cost situations ([72], p. 177).

We used the Weka environment for machine learning and the statistical package Stata 16 to estimate the models and for all calculations in the study.

Code 1. Pseudocode showing the strategy for algorithm comparison.

```
// Performance evaluation
for 200 times
  Random division of the sample in 10 stratified
  folds;
  for folds f = 1 to 10
    for algorithms a = 1 to n
      estimate the model using algorithm 'a' and data
      from all folds except 'f'
      using the model and data from fold 'f' compute
      TPCC, TPR, FPR, TNR, FNR, AUC;

// Statistical tests
for algorithms a = 1 to n
  for algorithms b = 1 to n
    conduct paired samples t test AUC(a) versus
    AUC(b);
```

Table 3 Mean and standard deviation for each performance indicator computed for each algorithm

	TPCC	TPR	FPR	TNR	FNR
RandomForest	94.0663	0.9439	0.0624	0.9376	0.0561
	(6.7569)	(0.0920)	(0.1002)	(0.1002)	(0.0920)
SVM (SMO)	93.6949	0.9605	0.0865	0.9135	0.0395
	(6.9304)	(0.0759)	(0.1153)	(0.1153)	(0.0759)
Naïve Bayes	91.9064	0.8935	0.0551	0.9449	0.1065
	(7.5032)	(0.1215)	(0.0897)	(0.0897)	(0.1215)
AdaBoost	88.8487	0.9116	0.1345	0.8655	0.0884
	(8.6396)	(0.1175)	(0.1422)	(0.1422)	(0.1175)
Multilayer Perceptron	87.7853	0.8844	0.1285	0.8715	0.1156
	(8.9901)	(0.1282)	(0.1388)	(0.1388)	(0.1282)
PART	87.6420	0.8919	0.1389	0.8611	0.1081
	(9.3924)	(0.1294)	(0.1455)	(0.1455)	(0.1294)
KNN	86.9010	0.9644	0.2264	0.7736	0.0356
	(9.0013)	(0.0730)	(0.1625)	(0.1625)	(0.0730)
J48	85.9846	0.8678	0.1479	0.8521	0.1322
	(8.7438)	(0.1329)	(0.1443)	(0.1443)	(0.1329)
Logit	85.2484	0.8414	0.1361	0.8639	0.1586
	(10.2138)	(0.1472)	(0.1487)	(0.1487)	(0.1472)

5 Results

Table 3 shows the statistics derived from the confusion matrix: TPCC, TPR, FPR, TNR, and FNR, for each of the different algorithms. In the table, algorithms are ranked according to TPCC and in each cell. The mean for the statistic is in the upper part of each cell, and the standard deviation is below (in parentheses).

Random Forests achieved the highest TPCC, as well as good values for the rest of statistics. However, the results produced by the SVM approach and, to a certain extent, by Naïve Bayes, were close. It is worth noting that for TPR, which is the ability to correctly identify ASD cases, SVM was, on average, better than Random Forests. Other models which had been used in related prior research produced worse results. Finally, all the machine learning models outperformed logistic regression, which we computed as a baseline approach.

Nevertheless, as indicated above, for an accurate comparison of the performance of the different algorithms when an accurate estimation of the cost matrix is not available, we conducted tests on the AUC estimations. As the tenfold estimation procedure implies that the same subsamples are used to compute the statistics for all the algorithms, we used a paired t-test to make comparisons between each possible pair of algorithms. Table 4 shows the mean and standard deviation for the AUCs for the different algorithms. Table 5 shows the results of the t-tests for the pairwise comparisons between the different algorithms. In both tables the algorithms are ranked according to the mean value of the AUCs.

Random Forests had the highest mean AUC and the lowest standard deviation, and statistically outperformed most of the algorithms. It is also worth noting that PART, which gave good results in previous studies, was the worst algorithm using the AUC-based metric. The results indicate that although in general the best performing algorithms were the same as in the previous studies summarized in Table 1, there were

Table 4 Mean value and standard deviation of AUC

	AUC mean	AUC Std. Dev
Rnd. Forest	0.9854	0.029
Naive Bayes	0.9663	0.054
KNN	0.9633	0.046
Ada Boost	0.9567	0.055
ANN	0.9453	0.064
SVM	0.9370	0.069
Logit	0.9122	0.083
J48	0.8947	0.094
PART	0.8854	0.096

Table 5 Results of the t-tests for the pairwise comparisons between algorithms

	Naive Bayes	KNN	Ada Boost	ANN	SVM	Logit	J48	PART
J48								-
Logit							-	-
SVM						-	-	-
ANN					-	-	-	*
Ada Boost				-	-	-	**	**
KNN		-		-	-	*	**	**
Naïve Bayes	-	-	-	-	-	*	**	**
Rnd., Forest	-	-	**	**	**	***	***	***

***The difference of means is significant at the 1% level

**The difference of means is significant at the 5% level

*The difference of means is significant at the 10% level

-The difference of means is not significant at the 10% level

some differences. This may be because of the size of the sample, but is more likely because those previous studies considered wider age ranges than our target sample, including older subjects for whom the symptoms of ASD are more evident than in subjects under 24 months old.

6 Robustness checks

Aside from the techniques evaluated above, we also estimated additional models used by previous researchers. These included Classification and Regression trees (CART) [73] and Hidden Markov Models (HMM) [74]. For all statistics, the results were lower than from the techniques indicated above, hence we do not include these results.

In addition, we also conducted an alternative comparison strategy consisting of keeping aside a completely withheld test set (20% of the original sample, randomly selected). This subsample was never a part of the training set. With the remaining 80%, and in order to be able to conduct t tests, we generated a number (n = 200) of artificial training subsamples. This was done by following a bootstrapping approach—randomly selecting with replacement several cases to form a sample the same size as

Table 6 Results of the t-tests for the pairwise comparisons between algorithms

Category	Video	Selected variables	Variables per category	Average variables per category
Social engagement	1	6	12	4
	2	4		
	3	2		
Social information gathering gaze exchanges	4	7	13	4,3
	5	1		
	6	5		
Gaze and deictic tracking	7	4	12	3
	8	5		
	9	1		
	10	2		

the original. With each of the artificial subsamples, we estimated the different ML models and computed the AUC statistic using the test set. This allowed us to make pairwise comparisons between algorithms using the paired samples t-test.[3] The results were qualitatively the same as the original strategy, confirming the main findings, the outstanding performance of Random Forests, and the relatively low classification power of PAGRT and J48.

Furthermore, we repeated the analysis, removing children over 18 months old from the sample. This subsample contained 50 ASD and 50 non-ASD children. The results from this subsample were qualitatively similar to those reported in Section 5. Finally, we estimated all the machine learning models and conducted the subsequent tests with the subset of 13 variables which were present in every fold of the tenfold estimation procedure. The statistics showed poorer performance from the algorithms, but the rankings were like those reported.

7 Discussion

7.1 Influence of video category

As mentioned in Section 4.3 (Stimuli and variables) the short movies used in the tests fell into three different categories: (i) social engagement, (ii) social information gathering gaze exchanges, and (iii) gaze and deictic tracking.

One of the advantages of our approach is that using a filtering process before applying the ML algorithms explained in Section 4.5 allows identification of the videos that provide more variables to the final set of indicators. We analyzed whether the distribution of the selected variables suggested a more determinant role for any of the three categories in the classification process. As Table 6 shows, there seems to be no influence.

---

[3] This procedure was implemented using Python code written by the authors, which is available from upon request.
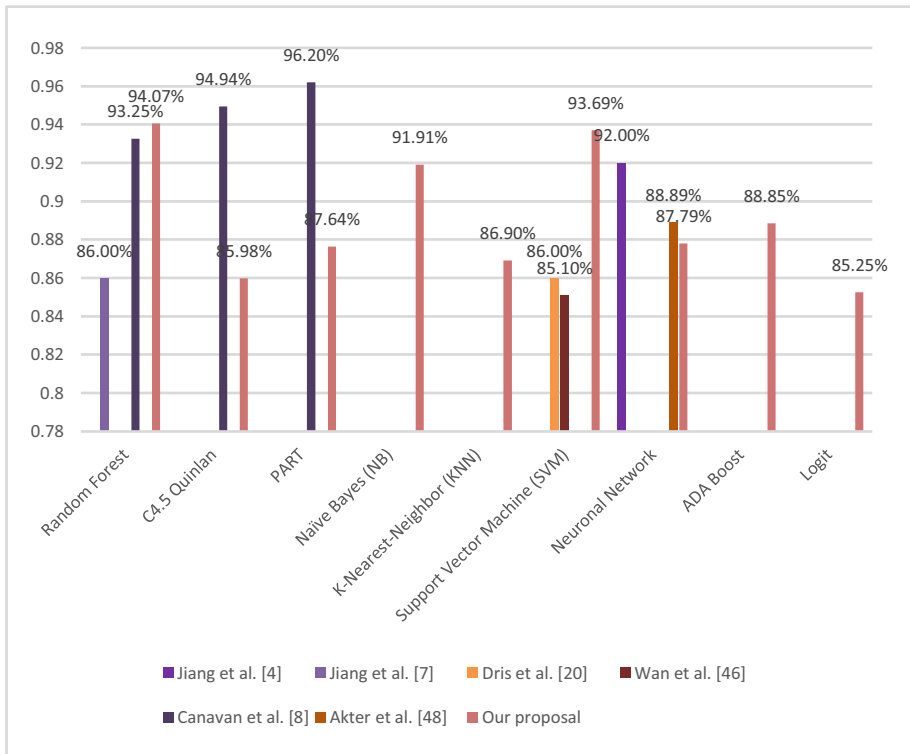
Fig. 6 Accuracy comparison with previous studies

Social engagement provided 12 variables, social information 13, and gaze and deictic tracking 12. This suggests that the influence of each video is more related to its design, rather than to the category of stimuli it belongs to. This can be explained by the results from Guillen et al. [6], discussed in Section 3.1, suggesting that the different behavior between ASD and TD subjects is strongly related to the context of the stimuli. Additionally, in future development, the videos contributing fewer variables could be removed, making for a shorter evaluation time for each subject.

7.2 Comparison with related work

Only a small proportion of previous studies considered the target age range—between 12 and 24 months (see Table 1). Most only studied children over the age of three, and some only looked at over-sixes, when the linguistic structures are completely established in normal development. Autism-related behavior becomes increasingly visible and detectable at this age.

Figure 6 compares past studies' ML-based classifiers to the various approaches examined in our study using this measure. It is worth noting that just one of the studies in Section 3 produced slightly better results.

Canavan et al. [9], achieved 96.2% accuracy employing PART (compared to our 94.07% using Random Forest). It is intriguing to note how Random Forest outperformed C4.5 and PART in our case, while these two algorithms produced better results in Canavan et al. This might be due to methodological differences. First of all they used the National Database for Autism Research (NDAR) Dataset [75] rather than a collection of ad-hoc stimulation films. Secondly, their approach also included demographic characteristics as categorization features. However, as previously indicated in Section 3, the most fundamental reason is that their sample comprised people ranging from two to more than sixty years old. This age group exhibits more obvious ASD-related behaviors than the neonates and toddlers tested in our study. Such a wide age range may produce a bias in the results when focusing on infants and toddlers.

The study by Alie et al. [32] is another of the few where the sample included individuals under 24 months old. As we saw in Section 3, they used pattern recognition algorithms in children around 6 months of age. Their Variable-order Markov Models achieved an accuracy of 93.75% with a sample of 32 individuals (26 non-ASD). This result is close to those from Canavan et al. [9] and to our results. However, the main disadvantage of their approach is that it cannot be automated since it requires parents' involvement and a manual analysis of the videos used as stimuli.

8 Summary and conclusions

The major issue our study addressed was determining how to lower the screening age from its present average of over 36 months to margins closer to the onset of ASD-related behavior (about 12 months).

To that end, we used a sample of babies and toddlers from an association for parents of ASD children. We used a matched pairs design to produce the overall sample, in which each ASD child was paired with a TD child of the same age. Many indications from gaze analysis were studied as independent variables using an eye-tracking device. The children were shown a series of films meant to provoke a reaction from them, producing a variety of metrics capturing social engagement, social information gathering gaze exchanges, and gaze following.

Those latter two categories have not been used in previous studies. We tested the performance of a number of ML methods, namely Random Forests, Naive Bayes, KNN, AdaBoost, ANN, SVM, J48, and PART, since they have already been used in prior research or are popular in classification literature. As a baseline approach, we also used logistic regression. The results show that several algorithms (Random Forests, ANN, Naive Bayes, and AdaBoost) produced good results (AUC better than 95%).

Random Forests outperformed the bulk of the systems tested. It is also worth mentioning that numerous methodologies that had produced excellent results in previous studies, such as PART, did not perform well with our sample. Furthermore, the robustness checks we conducted indicated that the results were qualitatively the same when only children under 18 months old were included.

This research presents a notable departure from prior work in several keyways. Some of the most notable contributions are:

1. Previous studies typically addressed limitations that necessitated manual intervention or expert knowledge, which is not the case in our automated approach.
2. Past research often involved subjects across a wide age range, which can compromise the accuracy of findings when establishing early-warning systems for autism spectrum disorder (ASD). This study is based exclusively on babies and toddlers.
3. Prior work did not conduct a comprehensive review of the various categorization algorithms available.
4. We conducted an analysis of test movies to capture social engagement, social information gathering gaze exchanges, and gaze following in babies and toddlers that is not included in previous studies. These findings can serve as valuable features for training machine learning algorithms.
5. Through a comparative analysis of multiple machine learning algorithms, we identified several that exhibited exceptional performance in diagnosing ASD in toddlers using gaze analysis.
6. Consequently, we have demonstrated the feasibility of an exceedingly effective (94.07%) automated eye-tracking-based early screening method for ASD in infants. This breakthrough could potentially facilitate early interventions and improve outcomes for children affected by ASD.

9 Limitations and future work

Despite the excellent results obtained thus far, we feel there is still room for improvement in the screening procedure. The method by Oliveira et al. [48], which is detailed in Section 3.2, is very relevant to our research.

We hypothesize that using automatically detected VAMs rather than expert-determined AOIs might improve the accuracy of the classification system.

In addition to this gaze-based approach, there are other ASD behaviors in toddlers that could be detected and processed automatically (for example, those related to motor response to a specific stimulus), providing another way to support early screening without the intervention of clinical experts.

Furthermore, it is notable that while the proposed approach is theoretically promising in terms of screening performance, our findings are based on the use of an eye tracker device, which is an expensive resource. Some previous studies, such as Bovery et al. [7], have yielded promising results using low-cost methods for monitoring gaze. We plan to investigate the feasibility of adapting the procedure to this less expensive strategy in combination with automatic VAM detection.

One notable limitation of our study is that it did not consider the potential impact of cultural or linguistic differences on the results, as the sample was drawn from a single country (Spain) and the videos used in the study were in Spanish. Although the video design was intended to be based on universal visual elements (commonly used objects and situations), it was not possible to assess the impact of using a specific language on subsequent work when adapting to other languages and cultures.

Finally, a functional prototype of the autism spectrum disorder screening application has been implemented, as depicted in Fig. 7. This application is designed for use
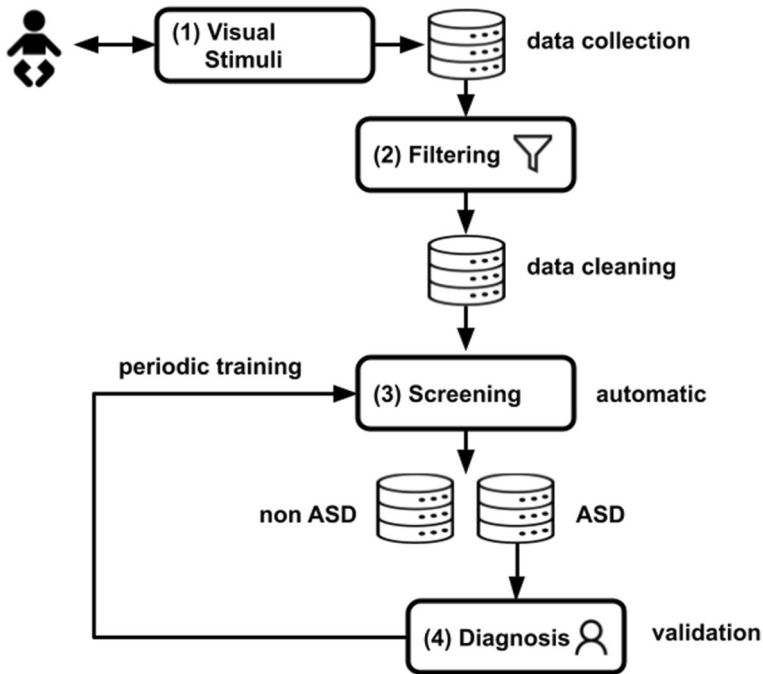
Fig. 7 Design of the automatic ASD screening prototype

by non-specialized personnel in healthcare centers during routine pediatric check-ups for infants aged between 12 and 24 months. Once again, the infants watch short movies (Step 1), but this time both data filtering and screening (Step 3) are performed automatically, with the screening using a Random Forest implementation. Infants classified as positive for ASD are referred to early intervention clinical services to initiate speech therapy treatment. It is important to note that the definitive diagnosis of ASD is made by a specialist (Step 4) when the child is older, typically around 3 years old, when language structures have further developed. Confirmed diagnoses of ASD through this method would be incorporated into the positive sample database, allowing for periodic training of the classification model to increase its accuracy as the sample size gradually expands. Funding is currently being sought for cloud application deployment and serial building of a less expensive version of the baby-friendly infrastructure outlined in Section 4.2.

Appendix

Table 7 Selected variables, descriptive statistics and significance of the t-test

Variable	Description	ASD = 0			ASD = 1		
		Mean	Std. Dev	Max	Mean	Std. Dev	Max
vid1_react_time**	Reaction time of the individual when shown video 1	0.4329	0.4458	1.62	0.2985	0.3116	1.51
vid1_num_fix_mouth***	Number of fixations in the mouth of the person appearing in video 1	8.2459	4.0359	19.00	4.5737	3.9262	13.00
vid1_num_fix_balloons***	Number of fixations in the balloons appearing in video 1	10.6557	4.0942	20.00	14.1475	8.2033	39.00
vid1_num_fix_eyes***	Number of fixations in the eyes of the person appearing in video 1	7.9508	6.5049	29.00	4.5737	4.9242	16.00
vid1_time_fix_balloons***	Time of fixation in the balloons appearing in video 1	2.6601	1.2228	4.88	4.3391	2.6862	12.68
vid1_time_fix_eyes***	Time of fixation in the eyes of the person appearing in video 1	4.3736	3.8032	13.99	2.6888	3.3012	15.23
vid2_num_fix_eyes**	Number of fixations in the eyes of the person appearing in video 2	4.7049	4.8039	20.00	3.1639	4.2236	14.00
vid2_num_fix_doll*	Number of fixations in the doll appearing in video 2	8.1639	4.8688	20.00	9.6885	6.7244	25.00
vid2_time_fix_eyes**	Time of fixation in the eyes of the person appearing in video 2	1.6844	2.0461	9.32	1.0103	1.4256	4.87
vid2_time_fix_doll	Time of fixation in the doll appearing in video 2	4.1498	2.6731	10.36	4.8536	3.8557	12.87
vid3_num_fix_wheel***	Number of fixations in the wheel appearing in video 3	7.6721	5.1046	19.00	14.4590	9.7100	38.00
vid3_time_fix_mouth***	Time of fixation in the mouth of the person appearing in video 3	2.5970	2.3425	10.13	1.5790	2.2492	12.08
vid4_num_fix_mouth***	Number of fixations in the mouth of the person appearing in video 4	7.9344	4.6435	20.00	3.4262	4.4998	18.00
vid4_num_fix_earrings***	Number of fixations in the earrings of the person appearing in video 4	1.0000	1.2247	5.00	0.1475	0.3575	1.00
vid4_num_fix_number***	Number of fixations in the number appearing in video 4	9.3442	10.8288	40.00	22.4262	11.4709	45.00
vid4_time_fix_mouth***	Time of fixation in the mouth of the person appearing in video 4	4.9780	3.7517	13.74	1.4650	2.3425	10.22
vid4_time_fix_earrings***	Time of fixation in the earrings of the person appearing in video 4	0.4363	0.6196	2.43	0.1290	0.5284	3.96
vid4_time_fix_number***	Time of fixation in the number appearing in video 4	3.9529	4.6948	19.03	11.9231	5.7480	19.63
vid4_num_ex_eyesmouth***	Number of exchanges between the eyes and the mouth of the person in video 4	0.9508	1.4767	6.00	0.0327	0.1795	1.00
vid5_num_ex_eeyesmouth***	Number of exchanges between the eyes and the mouth of the person in video 5	1.0327	1.2775	5.00	0.2950	0.8029	3.00
vid6_num_fix_eyes***	Time of fixation in the eyes of the person appearing in video 6	5.5737	6.3599	31.00	1.0655	2.0564	13.00
vid6_time_fix_mouth***	Time of fixation in the mouth of the person appearing in video 6	2.0503	2.0313	7.87	0.8011	0.9412	4.17
vid6_time_fix_eyes***	Time of fixation in the eyes of the person appearing in video 6	1.7424	2.3147	10.41	0.2380	0.5309	3.67

Table 7 (continued)

Variable	Description	ASD=0			ASD=1		
		Mean	Std. Dev	Max	Mean	Std. Dev	Max
vid6_num_ex_eyesmouth***	Number of exchanges between the eyes and the mouth of the person in video 6	1.2459	2.0054	11.00	0.0819	0.2765	1.00
vid6_num_ex_eeyesgenericobject***	Number of exchanges between the eyes of the person and the generic object in video 6	3.6065	3.5556	13.00	0.2131	0.6354	4.00
vid7_react_time	Reaction time of the individual when shown video 7	0.6791	0.6314	1.87	0.5996	0.5138	2.17
vid7_num_ex_eyesmouth***	Number of exchanges between the eyes and the mouth of the person in video 7	1.3114	1.6787	8.00	0.3114	0.6202	3.00
vid7_num_ex_eyescartoon	Number of exchanges between the eyes of the person and the cartoon in video 7	1.8196	2.4393	10.00	1.4918	1.5771	5.00
vid7_num_ex_fingercartoon***	Number of exchanges between the finger of the person and the cartoon in video 7	0.9508	1.7835	7.00	0.1475	0.4410	2.00
vid8_num_fix_empty3	Number of fixations in the 3rd empty space appearing in video 8	0.0163	0.1280	1.00	0.0819	0.3783	2.00
vid8_time_fix_error**	Time of erroneous fixation in video 8	0.7983	1.4039	5.39	0.3242	0.7705	3.24
vid8_num_ex_eyesfinger***	Number of exchanges between the eyes and the finger of the person in video 8	2.2622	2.0728	7.00	1.1967	1.8869	7.00
vid8_num_ex_eyesmouth***	Number of exchanges between the eyes and the mouth of the person in video 8	2.2786	2.3247	8.00	1.1475	1.6313	9.00
vid8_num_ex_mouthcartoon**	Number of exchanges between the mouth of the person and the cartoon in video 8	0.9016	1.5885	5.00	0.4262	0.8842	5.00
vid9_time_fix_empty2**	Time of fixation in the 2nd empty space appearing in video 9	0.0036	0.0197	0.11	0.0629	0.2002	1.02
vid10_time_fix_dog***	Time of fixation in the in the dog appearing in video 10	2.9409	2.1479	8.91	1.9760	1.7978	9.05
vid10_num_ex_mouthdog**	Number of exchanges between the mouth of the person and the dog in video 10	0.8196	1.3229	5.00	0.3278	0.8310	4.00

***The difference of means is significant at the 1% level

**The difference of means is significant at the 5% level

*The difference of means is significant at the 10% level

Data availability The datasets produced and analysed during the study are not publicly available due to the nature of the information, but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Elsabbagh M et al (2012) Global prevalence of autism and other pervasive developmental disorders. Autism Res 5(3):160–179
2. Saemundsen E, Magnússon P, Georgsdóttir I, Egilsson E, Rafnsson V (2013) Prevalence of autism spectrum disorders in an Icelandic birth cohort. BMJ Open 3(6):e002748
3. Buescher AVS, Cidav Z, Knapp M, Mandell DS (2014) Costs of autism spectrum disorders in the United Kingdom and the United States. JAMA Pediatr 168(8):721
4. Jiang M, Zhao Q (2017) "Learning visual attention to identify people with autism spectrum disorder," Proc EEE IntConf Comput Vision 2017:3287–3296
5. Takarae Y, Luna B, Minshew NJ, Sweeney JA (2008) Patterns of visual sensory and sensorimotor abnormalities in autism vary in relation to history of early language delay. J Int Neuropsychol Soc 14(6):980–989
6. Guillon Q, Hadjikhani N, Baduel S, Rogé B (2014) Changes in the focus of attention across time in individuals with autism: the effect of a dual-stream paradigm Jason. Neurosci Biobehav Rev 42:279–297
7. Bovery M, Dawson G, Hashemi J, Sapiro G (2021) A scalable off-the-shelf framework for measuring patterns of attention in young children and its application in autism spectrum disorder. IEEE Trans Affect Comput 12(3):722–731
8. Jiang M, Francis SM, Srishyla D, Conelea C, Zhao Q, Jacob S (2019) Classifying individuals with ASD through facial emotion recognition and eye-tracking. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, pp 6063–6068. https://doi.org/10.1109/EMBC.2019.8857005
9. Canavan S, Chen M, Chen S, Valdez R, Yaeger M, Lin H, Yin L (2018) Combining gaze and demographic feature descriptors for autism classification. In: Proceedings - International Conference on Image Processing, ICIP, 2017-Septe, pp 3750–3754. https://doi.org/10.1109/ICIP.2017.8296983
10. Drimalla H et al (2019) Detecting autism by analyzing a simulated social interaction. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G (eds) Machine learning and knowledge discovery in databases. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_12

11. Jones W, Carr K, Klin A (2008) Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. Arch Gen Psychiatry 65(8):946–954

12 American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders. American Psychiatric Association

13. Lord C, Risi S, DiLavore PS, Shulman C, Thurm A, Pickles A (2006) Autism from 2 to 9 years of age. Arch Gen Psychiatry 63(6):694

14. Lipkin PH (2017) Narrowing the diagnostic gap: Autism over 30 years. Infectiuous Diseases in Children. Issue 1

15. Dris AB, Alsalman A, Al-Wabil A, Aldosari M (2019) Intelligent gaze-based screening system for autism. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp 1–5

16. Osterling J, Dawson G (1994) Early recognition of children with autism: A study of first birthday home videotapes. J Autism Dev Disord 24(3):247–257

17. De Giacomo A, Fombonne E (1998) Parental recognition of developmental abnormalities in autism. Eur Child Adolesc Psychiatry 7(3):131–136

18. Øien RA, Vivanti G, Robins DL (2021) Editorial S.I: Early identification in autism spectrum disorders: The present and future, and advances in early identification. J Autism Dev Disord 51(3):763–768. https://doi.org/10.1007/s10803-020-04860-2

19. Kleinman JM et al (2008) The modified checklist for autism in toddlers: a follow-up study investigating the early detection of autism spectrum disorders. J Autism Dev Disord 38(5):827–839

20. Guthrie W, Wallis K, Bennett A, Brooks E, Dudley J, Gerdes M, Pandey J, Levy SE, Schultz RT, Miller JS (2019) Accuracy of autism screening in a large pediatric network. Pediatrics 144(4):e20183963. https://doi.org/10.1542/peds.2018-3963

21. Øien RA, Schjølberg S, Volkmar FR, Shic F, Cicchetti DV, Nordahl-Hansen A, Stenberg N, Hornig M, Havdahl A, Øyen AS, Ventola P, Susser ES, Eisemann MR, Chawarska K (2018) Clinical features of children with autism who passed 18-month screening. Pediatrics 141(6):e20173596. https://doi.org/10.1542/peds.2017-3596

22. Carette R, Elbattah M, Dequen G, Guérin J, Cilia F (2018) Visualization of eye-tracking patterns in autism spectrum disorder: Method and Dataset. In: 2018 Thirteenth International Conference on Digital Information Management (ICDIM), pp 248–253

23. 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 (2020) Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging 5(8):759–769. https://doi.org/10.1016/j.bpsc.2019.11.015

24. Jones W, Klin A (2013) Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. Nature 504(7480):427–431

25. Chawarska K, Macari S, Shic F (2013) Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. Biol Psychiatry 74(3):195–203

26. Wang S et al (2015) Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron 88(3):604–616

27. Pierce K, Conant D, Hazin R, Stoner R, Desmond J (2011) Preference for geometric patterns early in life as a risk factor for autism. Arch Gen Psychiatry 68(1):101–109

28. Pierce K, Marinero S, Hazin R, McKenna B, Barnes CC, Malige A (2016) Eye Tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. Biol Psychiatry 79(8):657–666

29. Frazier TW et al (2016) Development of an objective autism risk index using remote eye tracking. J Am Acad Child Adolesc Psychiatry 55(4):301–309

30. Vargas-Cuentas NI, Hidalgo D, Roman-Gonzalez A, Power M, Gilman RH, Zimic M (2016) Diagnosis of autism using an eye tracking system. In: GHTC 2016 - IEEE Global Humanitarian Technology Conference: Technology for the Benefit of Humanity, Conference Proceedings. Article 7857343 (GHTC 2016 - IEEE Global Humanitarian Technology Conference: Technology for the Benefit of Humanity, Conference Proceedings). Institute of Electrical and Electronics Engineers Inc, pp 624–627. https://doi.org/10.1109/GHTC.2016.7857343

31. Vargas-Cuentas NI, Roman-Gonzalez A, Gilman RH, Barrientos F, Ting J, Hidalgo D, Jensen K, Zimic M (2017) Developing an eye-tracking algorithm as a potential tool for early diagnosis of autism spectrum disorder in children. PloS one 12(11):e0188826. https://doi.org/10.1371/journal.pone.0188826

32. Alie D, Mahoor MH, Mattson WI, Anderson DR, Messinger DS (2011) Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using Markov models. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp 282–287

33. Tsang V (2018) Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders. Autism 22(2):161–170

34. Durand K, Gallay M, Seigneuric A, Robichon F, Baudouin JY (2007) The development of facial emotion recognition: The role of configural information. J Exp Child Psychol 97(1):14–27

35. Murias M et al (2018) Validation of eye-tracking measures of social attention as a potential biomarker for autism clinical trials. Autism Res 11(1):166–174

36. Del Coco M et al. (2017) "A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders," Proc 2017 IEEE IntConf Comput Vis Work ICCVW 2017 2018:1401–1407

37. Del Coco M et al (2018) Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot. IEEE Trans Cogn Dev Syst 10(4):993–1004

38. Kollias KF, Syriopoulou-Delli CK, Sarigiannidis P, Fragulis GF (2021) The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: a systematic review. Electron 10(23):2982

39. Hyde KK et al (2019) Applications of supervised machine learning in autism spectrum disorder research: a review. Rev J Autism Dev Disord 6(2):128–146 (Springer New York LLC)

40. Joudar SS, Albahri AS, Hamid RA (2022) Triage and priority-based healthcare diagnosis using artificial intelligence for autism spectrum disorder and gene contribution: a systematic review. Comput Biol Med 146:105553

41. Lawi A, Aziz F (2018) Comparison of classification algorithms of the autism Spectrum disorder diagnosis. In: 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT). IEEE, pp 218–222

42. Tyagi B, Mishra R, Bajpai N (2018) Machine learning techniques to predict autism spectrum disorder. In: 2018 ieee punecon. IEEE, pp 1–5

43. Wingfield B, Miller S, Yogarajah P, Kerr D, Gardiner B, Seneviratne S, Samarasinghe P, Coleman S (2020) A predictive model for paediatric autism screening. Health Informatics J 26(4), pp 2538–2553. https://doi.org/10.1177/1460458219887823

44. Altay O, Ulaş M (2018) Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp 1–4

45. Sarabadani S, Schudlo LC, Samadani AA, Kushski A (2020) Physiological detection of affective states in children with autism spectrum disorder. IEEE Trans Affect Comput 11(4):588–600

46. Krishnappababu PR, Di Martino M, Chang Z, Perochon SP, Carpenter KL, Compton S, Sapiro G (2021) Exploring complexity of facial dynamics in autism spectrum disorder. In: IEEE Transactions on Affective Computing

47. Wan G, Kong X, Sun B, Yu S, Tu Y, Park J, Lang C, Koh M, Wei Z, Feng Z, Lin Y, Kong J (2019) Applying eye tracking to identify autism spectrum disorder in children. J Autism Dev Disord 49(1):209–215. https://doi.org/10.1007/s10803-018-3690-y

48. Oliveira JS et al (2021) Computer-aided autism diagnosis based on visual attention models using eye tracking. Sci Rep 11(1):1–11

49. Akter, T., Ali, M.H., Khan, M.I., Satu, M.S., & Moni, M.A. (2021). Machine Learning Model To Predict Autism Investigating Eye-Tracking Dataset. 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp 383–387

50. Shic F, Macari S, Chawarska K (2014) Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. Biol Psychiatry 75(3):231–237

51. Chawarska K, MacAri S, Shic F (2012) Context modulates attention to social scenes in toddlers with autism. J Child Psychol Psychiatry Allied Discip 53(8):903–913

52. Thorup E, Nyström P, Gredebäck G, Bölte S, Falck-Ytter T (2016) Altered gaze following during live interaction in infants at risk for autism: An eye tracking study. Mol Autism 7(1):12

53. Lord C et al (1989) Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. J Autism Dev Disord 19(2):185–212

54. Lézine I, Brunet O (1992) *Le Développement psychologique de la première enfance*. Editions et Applications Psychologiques

55. Hall MA (2003) Correlation-based feature selection for machine learning

56. Doshi M, Chaturvedi SK (2014) Correlation based feature selection (CFS) technique to predict student perfromance. Int J Comput Networks Commun 6(3):197–206

57. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers

58. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106

59. Frank E, Witten IH (1998) Generating accurate rule sets without global optimization. International Conference on Machine Learning
60. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML'96). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 148–156
61. Dietterich TG (2000) Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40(2):139–157
62. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1. IEEE. pp 278–282
63. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
64. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
65. Vapnik VN (2000) The Nature of Statistical Learning Theory. Springer, New York
66. Russell SJ, Norvig P (2010) Artificial intelligence a modern approach. Pearson
67. John GH, Langley P (1995) Estimating continuous distributions in bayesian classifiers. Conference on Uncertainty in Artificial Intelligence
68. Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer Series in Statistics
69. Kuncheva LI (2006) On the optimality of Naïve Bayes. Pattern Recognit Lett 27(7):830–837
70. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6(1):37–66
71. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K (2017) Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. J Mach Learn Res 18:1–5
72. Witten IH, Frank E, Hall MA, Pal CJ (n.d.) *Data mining : practical machine learning tools and techniques*
73. Breiman L (2017) Classification and regression trees. Routledge
74. Baum LE (1972) An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. Inequalities 3:1–8
75. Payakachat N, Tilford JM, Ungar WJ (2016) National database for autism research (NDAR): big data opportunities for health services research and health technology assessment. Pharmacoeconomics 34(2):127–138