

Deep Learning for Early Behavioral Screening of Autism Spectrum Disorder Applications Advances and Translational Challenges

Xiaodi Chen *

Henan Polytechnic University, Jiaozuo 454000, China

* Corresponding author: (Email: 212410010003@home.hpu.edu.cn)

Abstract: Deep learning is increasingly used to convert brief, observable behaviors into quantitative markers for autism spectrum disorder (ASD) screening. This narrative review synthesizes peer-reviewed work published from September 2021 to September 2026 on eye tracking, facial and body video, speech and language, response-to-name tasks, questionnaires, and multimodal digital phenotyping. Recent studies show a shift from static image classification toward standardized behavioral elicitation, temporal representation learning, and fusion of complementary signals. Prospective evidence is strongest for tablet-based computer vision and structured eye-tracking paradigms, where clinically relevant performance has been reported in young children; however, many studies remain limited by small or convenience samples, sample-level leakage, weak differential-diagnosis controls, and absent external validation. Apparent accuracy therefore depends as much on cohort construction and validation design as on architecture. Convolutional networks remain common for spatial features, while recurrent, temporal convolutional, and Transformer models increasingly represent behavior over time. Multimodal systems can improve coverage of both social-communication differences and restricted or repetitive behaviors, but they introduce missing-data, calibration, privacy, and interpretability challenges. The field is best positioned to deliver clinician-supported screening and triage rather than autonomous diagnosis. Progress requires subject-level evaluation, prospective multicenter cohorts, transparent reporting, uncertainty-aware outputs, fairness analysis, and alignment of model evidence with clinical criteria.

Keywords: Autism spectrum disorder, early screening, deep learning, digital phenotyping, eye tracking, computer vision, multimodal learning.

1. Introduction

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental condition defined clinically by persistent differences in social communication and interaction together with restricted or repetitive patterns of behavior, interests, or sensory response [1, 2]. Characteristics may be detectable in early childhood, yet identification is often delayed, and the distribution of diagnostic resources is uneven [1]. Standard care integrates developmental history, caregiver report, observation, and clinical judgment; no single behavioral or biological measurement is sufficient. Consequently, automated systems should be framed as screening or decision-support tools that can prioritize assessment, quantify behaviors, or complement questionnaires—not as replacements for multidisciplinary diagnosis.

The appeal of behavioral artificial intelligence is practical. Cameras, microphones, eye trackers, tablets, and wearable sensors can record how a child attends, moves, vocalizes, or responds during short tasks. Deep neural networks can learn spatial and temporal representations from these high-dimensional signals, reducing dependence on hand-engineered features. Recent reviews nevertheless report substantial heterogeneity in modalities, reference standards, validation schemes, and outcome definitions [3, 4]. Reported area under the receiver operating characteristic curve (AUC) values may look impressive while concealing spectrum bias, repeated frames from the same participant across splits, or controls drawn only from typically developing children.

This review focuses on deep learning applied to observable behavior for early ASD screening from September 2021

through September 2026. Neuroimaging and electroencephalography are considered only as contrasts because their cost, acquisition burden, and clinical pathway differ from scalable behavioral screening. The central argument is that progress is moving from isolated images toward structured elicitation and multimodal digital phenotyping, but clinical readiness is determined by evaluation rigor, calibration, and workflow fit rather than by nominal accuracy alone.

As shown in Figure 1, this review organizes the field along five stages: behavioral data, standardization, deep representation, multimodal fusion and inference, and study and clinical outputs and traces how each stage contributes to, or constrains, the translation of deep-learning models into real screening pathways.

2. Review Method

A targeted narrative search was conducted across PubMed, Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar using combinations of autism or ASD with early screening, detection, deep learning, eye tracking, gaze, facial expression, video, gesture, movement, speech, language, response to name, wearable sensing, digital behavior, and multimodal fusion. The primary window was September 2021 to September 2026. Backward and forward citation tracking from recent systematic reviews was used to identify influential primary studies and methodological guidance.

Eligible studies involved children or an explicit early-screening pathway, analyzed directly observable behavioral data, included a neural-network component, and reported

validation outcomes. Studies using only MRI, EEG, genetics, microbiome, or intervention-response prediction were excluded from the main synthesis. Traditional machine learning was retained when it provided an external-validation benchmark or formed the decision layer on top of deep behavioral features. Conference papers were used cautiously when they supplied unique technical evidence; unverified preprints were not treated as established results.

For each core study, the synthesis considered sample composition, age, behavioral task, acquisition device, model family, split unit, comparator, discrimination metrics, external validation, interpretability, and limitations. Evidence was interpreted with attention to subject-level independence, diagnostic-reference quality, clinically relevant comparison groups, missing data, and prospective deployment. Because protocols and outcomes were too heterogeneous for meaningful pooling, results were summarized narratively rather than meta-analyzed.

3. Behavioral Data for Early ASD Screening

3.1. Eye Tracking and Visual Attention

Atypical allocation of attention to faces, eyes, biological motion, and competing nonsocial stimuli is measurable early in development. Eye tracking therefore offers a relatively direct bridge between experimentally elicited behavior and model input. Scan paths can be encoded as coordinate sequences, fixation and saccade statistics, regions of interest, heat maps, or images. Deep approaches use convolutional networks on heat maps, recurrent or gated models on trajectories, and hybrid CNN–GRU systems that combine spatial and temporal information.

The strongest recent evidence emphasizes protocol design and external testing. In a retrospective development cohort of 449 children followed by a prospective cohort of 80, Wei et al. obtained an external AUC of 0.849 with a random-forest model and showed that social and nonsocial gaze features were complementary [5]. Although this was not a purely deep model, it is an important realism benchmark: external performance was lower than many small internal studies. Chetcuti et al. demonstrated the feasibility of a two-minute protocol in infants and young children, supporting short standardized tasks while emphasizing feasibility and concurrent validity rather than autonomous diagnosis [6]. Fernandez-Lanvin et al. proposed an automated eye-tracking pipeline for toddlers younger than 24 months, but its clinical generalizability remains constrained by cohort size and site dependence [7].

Deep temporal models have attempted to preserve information lost when trajectories are collapsed into summary features. Benabderrahmane et al. fused gaze maps, fixation-distance sequences, and fixation-duration sequences with a CNN–GRU–ANN architecture and reported 93.1% precision on a validation dataset [8]. Colonnese et al. used data attribution and distillation to remove influential noisy examples, reporting 94.35% accuracy while retaining 77% of the dataset [9]. Such results illustrate useful directions—temporal modeling and training-data auditing—but also show why test composition and independence must accompany a headline metric. Image conversion may exploit mature vision backbones, yet it can obscure timing, stimulus context, and the distinction between a participant-level phenotype and repeated image samples [10].

3.2. Facial Expression Body Movement and Video

Video provides low-cost access to multiple behaviors, including gaze direction, affect, head pose, social orientation, joint attention, gesture, and repetitive movement. The task is technically difficult because clinically meaningful events are sparse, children move unpredictably, faces are frequently occluded, and home environments vary. Frame-level classifiers can exploit background, camera, or identity cues instead of behavior. Robust pipelines therefore require participant-level splits, tracking across time, and validation on different sites or devices.

A notable multicenter study trained deep-learning models on videos of joint-attention behavior and evaluated both ASD detection and symptom-severity estimation [11]. By tying the input to a clinically interpretable task, the study moved beyond generic facial appearance. Wu et al. used approximately 2,000 short infant videos to detect gaze, positive affect, and vocalization before combining behavior summaries for ASD classification [12]. More recent work has analyzed responses to sensory stimuli [13], home videos [14], and structured behavioral assessments. In a cohort of 241 autistic children, an open-source algorithm detected heterogeneous stereotypical motor movements with high sensitivity but more modest precision, demonstrating that quantifying a core symptom is distinct from classifying diagnosis [15].

Static facial-image classification remains common because public image collections and transfer-learning models are convenient. However, facial morphology is not equivalent to behavior, scraped datasets can contain duplicates and demographic confounds, and some highly cited facial-analysis work has been retracted [16]. Accordingly, this review treats static-face performance as low-certainty evidence unless provenance, subject independence, and external validation are clear. Clinically meaningful video research should quantify elicited social responses or repetitive behaviors and should test whether the model distinguishes ASD from developmental or language delays rather than only from typical development.

3.3. Speech Language and Social Response

Speech and language contain information about prosody, timing, turn taking, lexical diversity, echolalia, and social reciprocity. Deep acoustic encoders, recurrent networks, and pretrained speech representations can reduce manual feature engineering, while language models can represent transcripts. Yet age, language, recording quality, and verbal ability strongly influence performance. A speech classifier may separate minimally verbal children from age-matched controls without learning ASD-specific evidence, so comparison groups and stratification are essential.

Structured social probes reduce some variability. A multimodal response-to-name study recruited 125 toddlers across ASD, developmental delay, and typical-development groups, automatically quantified response score and timing, and reported an AUC of 0.81 for computer-rated results versus 0.91 for human ratings [17]. The result is clinically informative precisely because the system did not outperform observers: automation can standardize measurement and increase access while retaining a human reference. Simulated-interaction work combining facial and vocal characteristics provides a broader multimodal precedent, although early

studies often involved adults and should not be generalized directly to toddlers [18].

Future speech systems should report language background, hearing status, verbal level, device, task, and recording conditions. Self-supervised encoders may help in small datasets, but pretraining corpora can introduce age and accent bias. The most defensible use is to combine acoustic and conversational features with other behavioral signals and to output uncertainty when speech is absent or technically unusable.

3.4. Questionnaires Digital Phenotyping and Multimodal Behavior

Questionnaires are scalable but depend on caregiver interpretation and may perform differently across settings and demographic groups. Digital phenotyping augments them with directly measured behavior. In the prospective multiclinic SenseToKnow study, a tablet presented short stimuli while the frontal camera captured gaze, facial expression, head movement, blink rate, and motor behavior. Among 475 children aged 17–36 months, the combined algorithm achieved AUC 0.90, sensitivity 87.8%, specificity 80.8%, negative predictive value 97.8%, and positive predictive value 40.6% [19]. The high negative predictive value supports triage, while the lower positive predictive

value illustrates why a positive screen requires diagnostic follow-up.

An evaluated AI-based medical device combined a caregiver questionnaire, analysis of two short home videos, and a HCP questionnaire, returning positive, negative, or indeterminate results [20]. This architecture reflects real workflow needs: not every case should receive a forced binary output. Privacy-preserved crowd annotation has also been explored for home video, but scalable annotation does not remove the need for consent, data minimization, and safeguards against reidentification [21, 22]. Digital comorbidity patterns in health records can complement behavioral screening, yet they represent healthcare utilization as much as phenotype and may reproduce access inequities [23].

Multimodal fusion is attractive because ASD criteria span social communication and restricted or repetitive behavior. Early fusion concatenates features, late fusion combines modality-specific predictions, and attention-based fusion learns context-dependent weights. In practice, missing modalities, asynchronous sampling, and different noise levels are central design problems. A clinically useful model should degrade gracefully, disclose which modalities drove the result, and avoid treating a missing channel as diagnostic evidence.

We summarized the seven representative ASD behavior screening studies discussed above, as shown in Table 1.

Table 1. Representative behavioral screening studies

Study	Data and cohort	Model or task	Validation and result
Perochon et al. [19]	Tablet video; 475 children, 17–36 months	Multifeature computer vision and ML fusion	Prospective multiclinic; AUC 0.90; sensitivity 87.8%; specificity 80.8%
Wei et al. [5]	Eye tracking; 449 development + 80 prospective	Gaze features; RF benchmark	External AUC 0.849
JAMA joint-attention study [11]	Multi-institution child videos	Deep video model	Internal testing; diagnosis and severity tasks
Zhang et al. [17]	Response to name; 125 toddlers	Multimodal automated behavior coding	Computer-rated AUC 0.81; human-rated AUC 0.91
Benabderrahmane et al. [8]	Gaze maps and temporal sequences	CNN–GRU–ANN fusion	Validation precision 93.1%; external transportability unresolved
Colonnese et al. [9]	Eye-tracking sensor data	DNN with TraIn distillation	Test accuracy 94.35%; retained 77% of data
Noyman et al. [15]	319 assessments from 241 autistic children	Open-source movement detector	SMM sensitivity 92.53%; precision 66.82%

Note. AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; GRU, gated recurrent unit; ML, machine learning; RF, random forest; SMM, stereotypical motor movement.

4. Deep Learning Advances

4.1. Spatial and Temporal Representation

Convolutional neural networks remain the default for faces, gaze maps, and individual video frames because transfer learning reduces the number of task-specific parameters. Their weakness is temporal fragmentation. Averaging frame predictions ignores event order, while random frame splits can place near-duplicates from one child in training and testing. Recurrent networks, gated recurrent units, temporal convolutional networks, and 3D convolutions preserve sequences and are better matched to response latency, gaze transitions, gestures, and repetitive movements. The model should operate on a time scale justified by the behavioral construct rather than on whichever clip length maximizes

cross-validation accuracy.

Transformers offer long-range attention and flexible fusion, but their parameter count and data appetite are problematic in pediatric cohorts. Pretraining and self-supervised learning can improve efficiency by learning from unlabeled video or audio, yet transfer does not guarantee clinical relevance. A representation pretrained on adult internet video may encode appearance or production style more strongly than child social behavior. Fine-tuning should therefore include frozen-backbone baselines, ablation studies, identity-confound checks, and calibration on the intended population.

4.2. Multimodal Fusion Interpretability and Uncertainty

Multimodal models can integrate complementary evidence, but fusion architecture should follow the clinical question. Early fusion is compact when modalities are synchronized and complete. Late fusion is easier to audit and can preserve modality-specific calibration. Cross-modal attention can model interactions—such as whether head orientation follows a name call—but is vulnerable to spurious correlations when

sample sizes are small. Privacy-preserving multimodal and federated learning has been demonstrated in ASD research, though much of that work uses neuroimaging rather than scalable behavior [24, 25].

Interpretability should identify evidence that a clinician can inspect: stimulus-linked gaze allocation, response latency, smile dynamics, turn-taking, or movement episodes. Saliency maps on a face are insufficient if they highlight background or identity. Example attribution, temporal localization, and counterfactual tests can reveal whether predictions change for plausible behavioral reasons. Equally important is uncertainty. Calibration curves, confidence intervals, and an indeterminate category are more useful than a single accuracy value. A screening tool should be optimized for its position in the pathway: high sensitivity and negative predictive value may be preferred for initial triage, whereas specialist prioritization may require different thresholds.

5. Evidence Quality and Clinical Readiness

The literature contains a recurring gap between technical discrimination and clinical utility. Small balanced datasets simplify learning but do not represent the low prevalence and heterogeneous referrals encountered in primary care. Accuracy is particularly misleading under class imbalance; sensitivity, specificity, predictive values, AUC, calibration, and decision-curve consequences should be reported together. The SenseToKnow results illustrate this distinction: good sensitivity and AUC coexisted with a positive predictive value of 40.6% [19]. That is not a failure; it defines a screening tool that can reduce missed cases while routing positives to assessment.

Leakage is a major risk. Multiple frames, gaze images, or augmented samples from one participant must remain in a single partition. Hyperparameter selection should be nested within development data, and the final test set should be untouched. External validation should change at least one meaningful dimension—site, device, geography, language, or time. The externally tested gaze model [5] and prospective multiclinic digital-phenotyping study [19] are therefore more informative than near-perfect results from public image collections. Recent reviews similarly identify unclear patient selection, index-test bias, and inconsistent reporting as barriers to translation [3, 4].

Clinical reference standards also matter. Models trained against a prior label can reproduce diagnostic practice, including historical inequities, rather than detect an invariant phenotype. Controls limited to typical development inflate apparent specificity because the difficult clinical differential includes language delay, intellectual disability, attention-deficit/hyperactivity disorder, anxiety, and sensory or motor differences. Studies should report who established the diagnosis, whether assessors were blinded, and how indeterminate cases were handled.

Fairness cannot be inferred from overall accuracy. Sex, race and ethnicity, socioeconomic access, language, skin tone, device quality, and cultural norms can alter both behavior and measurement. Subgroup sample sizes and uncertainty should be reported, and threshold performance should be compared without assuming that parity on one metric guarantees equity. The prospective tablet study reported similar sensitivity across sex and racial or ethnic subgroups, an encouraging but not definitive finding because subgroup prevalence and

precision still matter [19].

Privacy is unusually salient because child video and voice are identifiable. Collection should be minimized to the task, retention should be limited, and secondary use should be explicit. On-device feature extraction, secure enclaves, federated learning, and privacy-preserved annotation may reduce exposure, but model updates and embeddings can still leak information. Families and autistic stakeholders should participate in decisions about acceptable capture, explanation, false-positive burden, and data governance.

Reporting and bias-assessment standards provide a practical route forward. TRIPOD+AI calls for transparent reporting of prediction-model development and evaluation [26]. PROBAST+AI supports assessment of participants, predictors, outcomes, and analysis in models using regression or AI [27]. For prospective interventional evaluation, CONSORT-AI and SPIRIT-AI specify additional information about input data, human–AI interaction, and error analysis [28, 29]. Applying these frameworks would make ASD screening studies easier to compare and less vulnerable to optimistic interpretation.

6. Research Gaps and Future Directions

First, research should move from retrospective convenience samples to prospective, consecutive recruitment across primary care and specialist services. Protocols should predefine the intended-use population, comparator pathway, operating threshold, and action after a positive, negative, or indeterminate result. Multicenter validation must include children with non-ASD developmental concerns, because differential screening—not separation from typical development—is the clinically relevant task.

Second, behavioral elicitation should cover both diagnostic domains. Many systems concentrate on social attention and facial response, leaving repetitive behavior and sensory reactivity underrepresented. Combining short gaze, response-to-name, joint-attention, vocal, and movement probes may provide broader criterion coverage without making the encounter burdensome. Temporal models should preserve stimulus timing and output localized evidence that clinicians can review.

Third, models need explicit strategies for heterogeneity. Age-conditional heads, mixture-of-experts models, hierarchical outcomes, and uncertainty-aware abstention are more plausible than a universal binary classifier. Performance should be stratified by developmental level, language, sex, and comorbidity, while avoiding subgroup overfitting. Calibration transfer and lightweight recalibration should be tested when models move across devices or clinics.

Fourth, data efficiency should be improved through self-supervised pretraining, weak supervision, and carefully governed federated learning rather than uncontrolled augmentation. Synthetic data may support engineering tests but should not substitute for representative clinical evaluation. Public benchmarks should provide participant-level split rules, hidden external test sets, acquisition metadata, and explicit licensing. Code, preprocessing decisions, and model versions should be archived when consent permits.

Finally, evaluation should measure clinical effect. Useful endpoints include referral appropriateness, time to assessment, missed-case rate, burden on families, clinician time, cost, and equity—not only AUC. Human-factors studies should test

how clinicians interpret explanations and uncertainty, whether automation bias occurs, and how disagreements are resolved. Digital tools can be valuable even when they do not exceed expert performance if they standardize observation, extend access, or identify when specialist review is needed. Adjacent reviews of neuroimaging, general screening instruments, triage, and conventional machine learning [30–34] reinforce the same need for representative cohorts and

transparent validation, but these modalities were outside the present behavioral deep-learning focus.

As shown in Table 2, from participant leakage and spectral bias to missing modalities, dataset shift, poor interpretability, and privacy and fairness concerns, each recurring transformation challenge corresponds to a specific methodological approach that can be built in during the research design phase rather than patched up afterward.

Table 2. Translational challenges and recommended responses

Challenge	Why it matters	Recommended response
Participant leakage	Inflates performance through identity or near-duplicate frames	Split and bootstrap at participant level; document all augmentation
Spectrum bias	Typical controls do not represent referrals	Include developmental and psychiatric differentials prospectively
Missing modalities	Fusion may learn missingness as a label	Mask channels, test graceful degradation, allow indeterminate output
Dataset shift	Sites, devices, languages, and ages change	External multicenter validation and calibration monitoring
Poor interpretability	Saliency may not correspond to behavior	Localize stimulus-linked events and provide inspectable evidence
Privacy and fairness	Child media are identifiable; errors may be unequal	Data minimization, subgroup audits, stakeholder governance
Unclear clinical role	High AUC does not specify action	Define pathway, threshold, downstream referral, and net benefit

7. Conclusion

Deep learning has expanded early ASD screening from questionnaires and hand-coded observations toward quantitative behavioral phenotyping. The most credible advances use short, standardized tasks; preserve temporal information; combine complementary signals; and evaluate performance prospectively or externally. Eye tracking and tablet-based computer vision currently provide the clearest evidence for scalable screening, while video, speech, and multimodal systems broaden coverage of clinically meaningful behavior. Nevertheless, small samples, leakage, unrepresentative controls, and missing external validation remain common. The appropriate near-term goal is not autonomous diagnosis but calibrated, privacy-conscious decision support that helps clinicians recognize risk, document behavior, and allocate assessment resources. Prospective multicenter evaluation, transparent reporting, stakeholder governance, and measurement of real clinical benefit are necessary before widespread deployment.

References

- [1] World Health Organization. (2025, September 17). Autism [Fact sheet]. <https://www.who.int/health-topics/autism>
- [2] American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.).
- [3] Valizadeh, A., Moassefi, M., Nakhostin-Ansari, A., et al. (2024). Automated diagnosis of autism with artificial intelligence: state of the art. *Reviews in the Neurosciences*, 35(2), 141–163. <https://doi.org/10.1515/revneuro-2023-0050>
- [4] Huda, S., Khan, D. M., Masroor, K., et al. (2024). Advancements in automated diagnosis of autism spectrum disorder through deep learning and resting-state functional MRI biomarkers: a systematic review. *Cognitive Neurodynamics*, 18(6), 3585–3601. <https://doi.org/10.1007/s11571-024-10176-z>
- [5] Wei, Q., Dong, W., Yu, D., et al. (2024). Early identification of autism spectrum disorder based on machine learning with eye-tracking data. *Journal of Affective Disorders*. <https://doi.org/10.1016/j.jad.2024.04.049>
- [6] Chetcuti, L., Varcin, K. J., Boutrus, M., et al. (2024). Feasibility of a 2-minute eye-tracking protocol to support the early identification of autism. *Scientific Reports*, 14, 5117. <https://doi.org/10.1038/s41598-024-55643-z>
- [7] Fernandez-Lanvin, D., Gonzalez-Rodriguez, M., De-Andres, J., et al. (2024). Towards an automatic early screening system for autism spectrum disorder in toddlers based on eye-tracking. *Multimedia Tools and Applications*, 83, 55319–55350. <https://doi.org/10.1007/s11042-023-17694-8>
- [8] Benabderrahmane, B., Gharzouli, M., & Benlecheb, A. (2024). A novel multi-modal model to assist the diagnosis of autism spectrum disorder using eye-tracking data. *Health Information Science and Systems*, 12, 40. <https://doi.org/10.1007/s13755-024-00299-2>
- [9] Colonnese, F., Di Luzio, F., Rosato, A., & Panella, M. (2024). Enhancing autism detection through gaze analysis using eye tracking sensors and data attribution with distillation in deep neural networks. *Sensors*, 24, 7792. <https://doi.org/10.3390/s24237792>
- [10] Alsharif, N., Al-Adhaileh, M. H., Al-Yatama, A., et al. (2024). Utilizing deep learning models in an intelligent eye-tracking system for autism spectrum disorder diagnosis. *Frontiers in Medicine*, 11, 1436646. <https://doi.org/10.3389/fmed.2024.1436646>
- [11] Ko, C., Lim, J. H., Hong, J., et al. (2023). Development and validation of a joint attention-based deep learning system for detection and symptom severity assessment of autism spectrum disorder. *JAMA Network Open*, 6, e2318159. <https://doi.org/10.1001/jamanetworkopen.2023.18159>
- [12] Wu, C., Liaqat, S., Helvaci, H., et al. (2021). Machine learning based autism spectrum disorder detection from videos. *IEEE Healthcom*. <https://doi.org/10.1109/HEALTHCOM49281.2021.9398924>
- [13] Serna-Aguilera, M., et al. (2024). Video-based autism detection with deep learning. 2024 IEEE Green Technologies Conference. <https://doi.org/10.1109/GreenTech58819.2024.10520462>
- [14] Singh, A., Rawat, A., Laroia, M., et al. (2024). Autism spectrum disorder screening on home videos using deep learning. *International Journal of Image, Graphics and Signal Processing*, 16(4). <https://doi.org/10.5815/ijigsp.2024.04.08>

- [15] Noyman, I., et al. (2024). Automated analysis of stereotypical movements in videos of children with autism spectrum disorder. *JAMA Network Open*.
- [16] Frontiers Editorial Office. (2023). Retraction: Deep learning for autism diagnosis and facial analysis in children. *Frontiers in Computational Neuroscience*, 17, 1215827. <https://doi.org/10.3389/fncom.2023.1215827>
- [17] Zhang, J., et al. (2023). A multimodal machine learning system in early screening for toddlers with autism spectrum disorders based on the response to name. *Frontiers in Psychiatry*, 14, 1039293. <https://doi.org/10.3389/fpsy.2023.1039293>
- [18] Drimalla, H., Scheffer, T., Landwehr, N., et al. (2020). Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task. *npj Digital Medicine*, 3, 25. <https://doi.org/10.1038/s41746-020-0227-5>
- [19] Perochon, S., Di Martino, J. M., Carpenter, K. L. H., et al. (2023). Early detection of autism using digital behavioral phenotyping. *Nature Medicine*, 29, 2489–2497. <https://doi.org/10.1038/s41591-023-02574-3>
- [20] Taraman, S., et al. (2022). Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. *npj Digital Medicine*, 5. <https://doi.org/10.1038/s41746-022-00598-6>
- [21] Washington, P., et al. (2022). Crowd annotations can approximate clinical autism impressions from short home videos with privacy protections. *Intelligence-Based Medicine*, 6, 100056. <https://doi.org/10.1016/j.ibmed.2022.100056>
- [22] Washington, P., et al. (2021). Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. *Scientific Reports*, 11.
- [23] Onishchenko, D., Huang, Y., van Horne, J., Smith, P. J., Msall, M. E., & Chattopadhyay, I. (2021). Reduced false positives in autism screening via digital biomarkers inferred from deep comorbidity patterns. *Science Advances*, 7.
- [24] Wang, H., Jing, H., Yang, J., et al. (2024). Identifying autism spectrum disorder from multi-modal data with privacy-preserving. *npj Mental Health Research*, 3, 15. <https://doi.org/10.1038/s44184-023-00050-x>
- [25] Chen, W., Yang, J., Sun, Z., et al. (2024). DeepASD: a deep adversarial-regularized graph learning method for ASD diagnosis with multimodal data. *Translational Psychiatry*, 14, 375. <https://doi.org/10.1038/s41398-024-02972-2>
- [26] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [27] Moons, K. G. M., Damen, J. A. A., Kaul, T., et al. (2025). PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388, e082505. <https://doi.org/10.1136/bmj-2024-082505>
- [28] Liu, X., Cruz Rivera, S., Moher, D., et al. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26, 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
- [29] Cruz Rivera, S., Liu, X., Chan, A. W., et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, 26, 1351–1363. <https://doi.org/10.1038/s41591-020-1037-7>
- [30] Santana, C. P., de Carvalho, E. A., Rodrigues, I. D., et al. (2022). Resting-state functional MRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Journal of Neuroscience Methods*, 367, 109454. <https://doi.org/10.1016/j.jneumeth.2021.109454>
- [31] Moridian, P., Ghassemi, N., Jafari, M., et al. (2022). Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: a review. *Frontiers in Molecular Neuroscience*, 15, 999605. <https://doi.org/10.3389/fnmol.2022.999605>
- [32] Thabtah, F., Spencer, R., Abdelhamid, N., et al. (2022). Autism screening: an unsupervised machine learning approach. *Health Information Science and Systems*, 10, 26. <https://doi.org/10.1007/s13755-022-00191-x>
- [33] Santos, C. L., Barreto II, Floriano, I., et al. (2024). Screening and diagnostic tools for autism spectrum disorder: systematic review and meta-analysis. *Clinics*, 79, 100323. <https://doi.org/10.1016/j.clinsp.2023.100323>
- [34] Joudar, S. S., Albahri, A. S., Hamid, R. A., et al. (2022). Triage and priority-based healthcare diagnosis using artificial intelligence for autism spectrum disorder and gene contribution: a systematic review. *Computers in Biology and Medicine*, 146, 105553. <https://doi.org/10.1016/j.compbiomed.2022.105553>