Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Algorithms to Clinics: Evaluating AI's Role in Pediatric Autism Diagnosis

Nikesh Adhikari

learnnepalnikesh@gmail.com

Coventry University https://orcid.org/0009-0005-4415-7258

Nava Raj Gautam

gautamn6@uni.coventry.ac.uk

Coventry University https://orcid.org/0009-0006-8347-4296

Systematic Review

Keywords: autism, artificial intelligence, machine learning, screening, diagnostic support, external validation, clinical utility, explainable artificial intelligence, neurodiversity, health equity

Posted Date: September 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11076952/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Artificial intelligence (AI) is increasingly proposed as a means of accelerating autism screening and supporting diagnostic assessment. Questionnaire, facial, vocal, video, eye-tracking, and multimodal systems sometimes report accuracy above 90%. Such estimates, however, often arise from small or curated datasets, internally derived labels, balanced case–control samples, or cross-validation conducted within a single dataset. Consequently, benchmark performance may substantially overstate clinical utility. This critical integrative review synthesizes evidence on AI for pediatric autism screening and diagnostic support, emphasizing the transition from model accuracy to safe, reproducible, and equitable use. The five records supplied for the review were appraised as seed sources and expanded through structured searches and backward and forward citation chaining through July 21, 2026. Peer-reviewed studies were evaluated by clinical role, reference standard, sample provenance, leakage controls, validation design, calibration, abstention, subgroup performance, explainability, reproducibility, and stakeholder involvement. Retracted and non-peer-reviewed records were not treated as evidence of efficacy. Internally evaluated questionnaire, image, voice, and behavioral models frequently reported high discrimination, but performance generally attenuated in prospective, geographically independent, or naturalistic testing. External or prospective evaluations produced area-under-the-curve estimates of approximately 0.73–0.90 rather than the near-perfect values found in some curated-dataset studies. Positive predictive value, clinical coverage, and false-referral burden varied with deployment prevalence and model abstention. Feature-attribution methods improved technical interpretability but rarely demonstrated better clinician decisions or family understanding. Representation gaps, incomplete reporting of calibration and uncertainty, inadequate external validation, and limited participation by autistic people constrained translation. Current evidence supports AI primarily as an adjunct for screening, triage, and structured decision support—not as an autonomous diagnostic authority. A five-stage clinical-readiness framework is proposed, spanning data validity, model validity, clinical validity, workflow utility, and equitable benefit.

1. Introduction

Autism is a lifelong neurodevelopmental difference associated with heterogeneous patterns of social communication, sensory processing, interests, behavior, and support needs. Clinical identification is not reducible to a single biomarker or questionnaire score. It ordinarily requires developmental history, direct observation, information from caregivers and other settings, consideration of co-occurring conditions, and professional judgment. The resulting process can be resource intensive, and access to appropriately trained clinicians is uneven. These conditions make autism an attractive target for computational screening and decision-support systems.

AI research in this domain has expanded from algorithms that shorten established questionnaires to systems that analyze home video, gaze, facial behavior, vocal characteristics, text, medical history, and combinations of these modalities. The promise is compelling: a scalable tool could identify children who need further assessment, prioritize referrals, structure observations, or help generalist clinicians use

limited specialist capacity more effectively. The language used to describe such systems, however, frequently moves too quickly from “classification” to “diagnosis.” A model that separates previously labeled groups in a research dataset has not necessarily demonstrated that it can diagnose an individual child in primary care.

This distinction matters because screening and diagnosis have different purposes, error costs, and evidentiary requirements. Screening identifies elevated likelihood and is expected to generate false positives that must be resolved by assessment. Diagnostic support contributes information to a clinician who integrates multiple sources. Autonomous diagnosis would make or effectively determine a clinical conclusion. Most current autism AI systems have evidence relevant to the first role, a smaller number have been evaluated for the second, and the evidence is not sufficient for the third (Erden et al., 2021; Megerian et al., 2022).

Published performance estimates can also be misleading when interpreted outside their study design. Accuracy depends on class balance and threshold choice. Sensitivity and specificity do not show whether predicted probabilities are calibrated. Positive and negative predictive values change with prevalence. Cross-validation within one dataset cannot establish transportability to a different clinic, language, culture, device, or referral pathway. Models can also inherit circularity when predictors incorporate items or scores closely related to the outcome label. These issues are especially consequential in autism, where the phenotype is heterogeneous and clinical labels can reflect differences in access, referral practices, masking, language, intellectual functioning, sex and gender, race, socioeconomic position, and local diagnostic norms.

Several recent reviews describe AI modalities for autism detection. Mohammadi et al. (2025) mapped facial, vocal, and textual approaches, Jia et al. (2025) reviewed AI screening methods, and Sun et al. (2025) meta-analyzed algorithms tracking social behavior. A broader public-health review subsequently examined AI across the autism-care continuum and health equity (Han, Weng, & Xu, 2026). The present review makes a different contribution. Rather than cataloging modalities or treating reported accuracy as the primary outcome, it examines what evidence is required to move from benchmark performance to clinical readiness.

Four recurring conflations organize the analysis: screening versus diagnosis; internal validation versus independent clinical validation; feature attribution versus clinically meaningful explanation; and discrimination performance versus equitable health benefit. The review asks: (a) What roles are current AI systems actually capable of supporting? (b) Why do high benchmark scores often attenuate under external or prospective evaluation? (c) Which methodological, ethical, and implementation gaps prevent clinical translation? and (d) What minimum evidence should precede deployment?

2. Methodology

2.1 Review design

A critical integrative review design was selected because the supplied material combined an empirical modeling study, a scoping review, a magazine feature, an undergraduate project report, and a subsequently retracted article. Treating these records as equivalent inputs to a conventional systematic review would obscure major differences in evidentiary status. The integrative design permitted synthesis of empirical performance studies, evidence reviews, clinical-translation analyses, and sociotechnical critiques while preserving explicit distinctions between them.

The review was designed to generate a clinically oriented argument rather than a pooled estimate across incompatible modalities. No meta-analysis was undertaken because populations, index tests, comparator groups, reference standards, outcome definitions, validation designs, and reported metrics were substantially heterogeneous.

2.2 Search and source expansion

The five user-supplied records were used as seed sources. Structured searches were then conducted through July 21, 2026, using PubMed, publisher databases, Crossref-indexed records, and targeted scholarly web searching. Backward and forward citation chaining was applied to the seed sources and the most clinically developed studies. Search concepts combined terms for autism, child or pediatric populations, artificial intelligence or machine learning, screening, diagnosis, questionnaire, home video, digital phenotyping, eye tracking, facial behavior, voice, explainability, external validation, fairness, and clinical utility.

Ten peer-reviewed sources were selected for focal expansion to provide methodological and temporal breadth: foundational questionnaire reduction, combined questionnaire/video screening, prospective home-video validation, a regulated multimodal medical device, prospective digital behavioral phenotyping, large-cohort medical-history prediction, multisite automated video analysis, meta-analytic evidence on social-behavior and eye-tracking systems, and critical analysis of automation in autism assessment. Additional reporting standards and sociotechnical scholarship were used to evaluate transparency, bias, and neurodiversity-related concerns.

2.3 Eligibility and exclusion criteria

Eligible focal sources were English-language peer-reviewed journal articles that: (a) evaluated an AI or machine-learning method for pediatric autism screening, case identification, or diagnostic support; (b) synthesized diagnostic or screening evidence across relevant AI modalities; or (c) directly addressed clinical, ethical, or sociotechnical translation of autism assessment AI. Studies of intervention personalization without a screening or diagnostic component, purely genetic or neuroimaging biomarker studies with no plausible near-term pediatric workflow, conference abstracts without sufficient methods, promotional materials, and duplicate reports were excluded from the focal synthesis.

Non-peer-reviewed material could inform context but not efficacy claims. Retracted work was excluded from all performance conclusions. It was retained only when the retraction itself provided evidence relevant to research integrity and reproducibility.

2.4 Data extraction and critical appraisal

For each focal empirical study, the review extracted intended clinical role, population and setting, modality, reference standard, sample size, class distribution, modeling approach, validation strategy, discrimination and classification metrics, handling of indeterminate outputs, subgroup analysis, explainability method, data and code availability, and declared conflicts of interest. Reviews were assessed for search breadth, study heterogeneity, external-validation findings, and limitations.

Because the evidence included development studies, diagnostic-accuracy studies, reviews, and conceptual analyses, a single numerical quality score was inappropriate. Appraisal domains were instead adapted from TRIPOD + AI and STARD-AI: dataset provenance, participant flow, reference-standard independence, separation of development and evaluation data, leakage controls, missing-data handling, calibration, uncertainty, external validation, subgroup performance, transparency, and clinical applicability (Collins et al., 2024; Sounderajah et al., 2025).

2.5 Synthesis and limitations of this review

Evidence was organized by clinical role and modality, followed by cross-cutting analysis of validation, prevalence, abstention, explainability, fairness, reproducibility, and stakeholder governance. This review is critical and integrative, not an exhaustive systematic review. Searches were targeted rather than designed to identify every algorithmic study, and no independent second reviewer conducted screening or extraction. The conclusions therefore concern patterns of evidence and translation rather than a definitive count of all published models. This limitation is made explicit to avoid a false claim of PRISMA-level comprehensiveness.

3. Status of the Supplied Sources

The five supplied records were not equivalent research articles. Their classification materially changed the synthesis.

Table 1

Evidentiary classification of the five supplied records

Source	Evidentiary status	Use in this review
Mohammadi et al. (2025)	Peer-reviewed scoping review of 23 studies involving facial, vocal, and textual methods	Included as a core modality map; high reported accuracies are interpreted in light of small, curated, and heterogeneous datasets
Itu et al. (2021)	Undergraduate project report hosted on ResearchGate; not an independently peer-reviewed clinical study	Used only to illustrate speculative assistive-device narratives; not used as evidence of safety, accuracy, or effectiveness
Jiang retraction notice (2025)	Retraction of the supplied Scientific Reports article	Original performance claims excluded; retraction used as a case of unverifiable data provenance, unavailable code, inadequate methods, and problematic generated content
Mertz (2024)	IEEE Pulse feature article rather than an original validation study	Used as contextual reporting on emerging technologies and commercial translation
Alsharif et al. (2025)	Peer-reviewed modeling study using public Saudi and Egyptian screening datasets	Included with substantial methodological caution regarding screening-versus-diagnosis ambiguity, possible circular predictors, cross-validation-only evaluation, and causal overinterpretation

Mohammadi et al. (2025) reported promising ranges across face, voice, and text analysis, including very high values in some facial-image studies. These ranges are useful for showing the field's technical breadth but should not be treated as pooled clinical diagnostic accuracy. The underlying studies used different tasks, populations, labels, and evaluation designs.

The ResearchGate report by Itu et al. (2021) proposed a device that would infer needs and risks from facial expressions and alert caregivers. It did not provide a clinical validation study capable of establishing those functions. Its inclusion without qualification would blur the difference between a design concept and evidence.

The supplied Scientific Reports article was retracted shortly after publication. The retraction notice stated that the framework was inadequately described, a generated figure contained nonsensical material, code was unavailable, and neither the claimed training nor validation dataset could be verified (Jiang, 2025). Its reported accuracy is therefore not cited as evidence. The episode is nevertheless relevant because it shows why model provenance, code, and traceable evaluation data are not optional reporting details.

Alsharif et al. (2025) reported 97% and 92% accuracy using K-nearest neighbors on Saudi and Egyptian datasets, respectively. Both estimates came from five-fold cross-validation rather than held-out geographic or prospective validation. More importantly, the Saudi analysis identified a "screening score" as its most important predictor. Unless the outcome label was independently established, such a

predictor risks reproducing the instrument or label used to define the target. The paper also moved from predictive associations and SHAP values to language about causal risk factors and treatment planning. These are not warranted inferences from a classification model.

4. Thematic Review

4.1 Clinical role: Screening is not diagnosis

The most important organizing question is not which algorithm performs best but what decision it is intended to support. A primary-care screener may be valuable if it reliably identifies children who need assessment, even though many positive screens are not ultimately diagnosed as autistic. A triage model may prioritize cases already referred. A diagnostic-support device may contribute structured observations to a clinician. These roles use different populations and thresholds and impose different costs for false-negative, false-positive, and indeterminate results.

Much of the literature evaluates retrospective classification between autistic and non-autistic groups. In balanced case–control data, accuracy can appear high because the prevalence is artificially close to 50%. In routine primary care, where prevalence among all screened children is much lower, the same sensitivity and specificity can yield a modest positive predictive value and a substantial referral burden. Conversely, a specialist referral population may include many children with language delay, attention-deficit/hyperactivity disorder, intellectual disability, anxiety, or other developmental presentations that make the classification task more difficult than distinguishing autism from typical development.

The language of “AI diagnosis” can therefore exceed the evidence. Bone et al. (2016) demonstrated that machine learning could reduce items in established instruments while preserving useful classification information. This is a contribution to assessment efficiency, not proof that the shortened algorithm independently performs the full diagnostic process. Erden et al. (2021) similarly argued that AI may add consistency and structure but cannot replace the interpretive and relational work through which clinicians understand development in context.

4.2 Questionnaire and minimal-feature models

Questionnaire-based models are attractive because they are inexpensive, scalable, and readily digitized. They also illustrate the risk of circularity. If a model is trained to predict a label derived partly from the same or closely related questionnaire items, it may learn the scoring logic of an existing instrument rather than discover an independent clinical signal.

Bone et al. (2016) used machine learning to improve the efficiency of screening and diagnostic instruments and explored fusion across instruments. The work showed that sparse behavior sets could retain considerable predictive information. Its translational value lies in reducing burden and identifying redundant items. Yet performance based on expert-research datasets does not automatically transfer to

unselected community screening, translated instruments, or settings in which follow-up questions are inconsistently administered.

Abbas et al. (2018) combined a short caregiver questionnaire with coded home video in a clinical sample of 162 children. The study showed that multimodal fusion and an inconclusive category could improve classification among cases for which the system generated a result. The authors appropriately acknowledged statistical limitations and the need for larger clinical studies. This pattern—strong results accompanied by incomplete coverage—would later become central to regulated-device evaluation.

Rajagopalan et al. (2024) developed a model using 28 medical and background features in 30,660 SPARK participants. The best internal model achieved an area under the receiver operating characteristic curve (AUROC) of 0.895. When tested on independent SPARK and Simons Simplex Collection cohorts totaling 14,790 participants, AUROC declined to 0.790. The attenuation does not invalidate the model; it makes the estimate more informative. It demonstrates both genuine signal and the cost of transport beyond the development data. Because the cohorts were research registries with highly uneven autism prevalence, further population-based and clinical-pathway validation remains necessary.

Alsharif et al. (2025) provide the opposing example. Cross-validated accuracy above 90% was presented as evidence for early diagnosis and clinical planning, despite reliance on public screening data, unclear independence of labels and predictors, and no prospective evaluation. High discrimination under these conditions should be treated as model-development evidence, not clinical validation.

4.3 Home video and digital behavioral phenotyping

Video systems aim to capture observable social and motor behavior outside specialist clinics. Their evolution illustrates increasing methodological realism. Tariq et al. (2018) developed sparse classifiers from tagged home videos. A five-feature model achieved an AUC of 0.92 in development and 0.89 in a prospectively collected independent set of 66 videos. A newly learned eight-feature model achieved an AUC of 0.93 on a held-out set but 0.86 in the prospective validation set. Human raters still tagged behaviors, so the workflow was mobile and distributed rather than fully automated. The study nevertheless represented an important advance over within-dataset cross-validation.

Megerian et al. (2022) evaluated a software medical device in a double-blinded, multisite prospective cohort of 425 children aged 18–72 months who had developmental concerns. The system combined a caregiver questionnaire, a healthcare-provider questionnaire, and two caregiver-recorded videos scored by trained analysts. Compared with specialist agreement, the positive predictive value was 80.8% and negative predictive value 98.3%. Sensitivity of 98.4% and specificity of 78.9% applied only to the 31.8% of children who received a determinate result. The system abstained for 68.2% of participants, many of whom had complex neurodevelopmental presentations. If the abstention mechanism were removed, sensitivity and specificity would have fallen sharply.

This study offers two important lessons. First, abstention can be a legitimate safety mechanism. Second, headline sensitivity without coverage is incomplete. A system may perform well precisely

because it refers difficult cases back to human specialists. That can still have clinical value if the determinate subset is large enough to reduce workload safely, but the relevant outcome is pathway efficiency and missed-case burden—not sensitivity alone. The study also involved trained video analysts and substantial company-related interests, emphasizing the need to evaluate the complete sociotechnical system and to replicate results independently.

Perochon et al. (2023) evaluated the SenseToKnow digital phenotyping application prospectively during pediatric well-child visits. Among 475 children aged 17–36 months, a model combining multiple app-derived phenotypes achieved an AUROC of 0.90, sensitivity of 87.8%, specificity of 80.8%, negative predictive value of 97.8%, and positive predictive value of 40.6%. Sensitivity was similar across reported sex, race, and ethnicity subgroups, although confidence intervals widened in smaller groups. The modest positive predictive value is not a failure: it reflects the function of a screener in a lower-prevalence setting. It also shows why AUROC and accuracy cannot by themselves communicate what clinicians and families will experience.

Kim et al. (2025) moved further toward automated naturalistic assessment. Three video tasks lasting under one minute each were collected from 510 children aged 18–48 months across nine South Korean hospitals. The ensemble model achieved AUROC 0.83 and accuracy 0.75. External testing on noisier recordings produced AUROC 0.73. These values are lower than many curated facial-image reports but more credible for translation because the study included multiple sites and variable home-recording conditions. The decline under noisy external data quantifies a practical distribution shift rather than an abstract limitation.

4.4 Facial behavior, gaze, voice, text, and multimodal systems

Mohammadi et al. (2025) identified 23 studies applying AI to facial, vocal, and textual data. Deep neural networks dominated facial-image analysis, while voice and text studies more often used conventional machine-learning approaches. Reported results were often high, but sample size, capture conditions, demographic diversity, and validation practices varied substantially. Static facial morphology deserves particular caution because it may encode camera, ancestry, age, expression, or dataset artifacts rather than an autism-specific phenotype.

Dynamic social behavior is more clinically plausible than decontextualized facial appearance, but it remains context dependent. Sun et al. (2025) reviewed 38 studies of AI-based social-behavior tracking and included seven in a meta-analysis. The estimated summary AUROC was 0.862, with a pooled diagnostic odds ratio of 15.917. The review found promising measurement of gaze, emotional expression, and joint attention, while emphasizing the need for validation in diverse populations. Only a minority of eligible studies provided sufficient data for quantitative synthesis, itself indicating inconsistent reporting.

Eye-tracking studies offer objective measurements but have frequently used small, controlled samples. Han, Yang, et al. (2026) synthesized 25 studies with 2,319 participants and reported pooled accuracy around 85%, with sensitivity and specificity around 86%. The review also identified high heterogeneity, small samples, and an absence of external validation. Eye tracking may therefore contribute a useful behavioral signal, but evidence does not support treating it as a stand-alone diagnostic marker across populations and devices.

Voice and language systems face comparable challenges. Vocal prosody, turn taking, word choice, and interaction patterns vary with age, language, intellectual and language development, anxiety, speech-motor differences, interlocutor behavior, and recording environment. A classifier that works in one language or structured task may fail when these contextual variables change. Multimodal fusion can reduce reliance on any single signal, but it also increases missingness, infrastructure requirements, privacy exposure, and the possibility that one shortcut feature dominates the model.

4.5 Explainability: Attribution is not understanding

Explainable AI is often presented as the solution to clinical trust. In autism research, SHAP, LIME, feature importance, and decision trees are used to show which variables influenced a prediction. These tools can help detect implausible dependencies and communicate model behavior. They do not, however, establish causation, validate a feature clinically, or prove that an explanation improves a decision.

Alsharif et al. (2025) illustrates the problem. Feature importance was used to discuss demographic and environmental “risk factors” and to motivate intervention implications. A high SHAP value means that a feature influenced a model prediction relative to a specified background distribution. It does not show that changing the feature would change autism likelihood, that the relation is causal, or that the feature should guide treatment. When the dominant predictor is a screening score, the explanation may simply reveal circularity.

Explanations must therefore be evaluated as interventions. Relevant questions include whether clinicians detect errors more accurately, whether families understand the system’s limits, whether explanations encourage appropriate override, and whether autistic people consider the categories and language respectful. A visually polished feature-importance plot is not evidence of meaningful interpretability. The Alsharif study itself acknowledged that clinician and educator user studies had not yet been conducted.

4.6 Clinical workflow and human–AI complementarity

The near-term value of AI lies in redistributing work rather than replacing judgment. Systems may automate questionnaire scoring, identify low-risk cases, structure observations, prioritize referrals, or flag recordings needing expert review. Their benefit should be measured at the pathway level: time to assessment, specialist hours saved, proportion of children receiving a determinate result, false reassurance, unnecessary referrals, family burden, and access across geographic and socioeconomic groups.

Megerian et al. (2022) suggests that selective automation can produce high predictive values for a subset of children while deferring complex cases. Perochon et al. (2023) suggests that a brief app could strengthen screening during routine visits. Kim et al. (2025) suggests that automated home-video analysis may support referral prioritization under realistic recording variability. None establishes that clinicians can be removed from the pathway. The safest design is a complementarity model: the system handles a clearly specified task, communicates uncertainty, and preserves human authority and accountability.

5. Critical Analysis

5.1 Why benchmark accuracy becomes inflated

Several design choices can inflate apparent performance. Small samples allow unstable estimates and wide but unreported uncertainty. Balanced case–control sampling makes accuracy easier to interpret mathematically but unrealistic clinically. Comparing autistic children only with typically developing controls excludes the diagnostically difficult middle: children with other developmental, language, behavioral, or psychiatric differences. Repeated observations from the same child can leak across training and test folds. Preprocessing or feature selection conducted before data splitting allows information from the test set to influence the model. Outcome-derived screening scores can encode the label directly.

Cross-validation reduces dependence on a single split but does not solve these problems. If every fold comes from the same source dataset, all folds share recruitment practices, devices, annotation conventions, and label processes. The 97% cross-validated accuracy reported by Alsharif et al. (2025) and the 0.73 AUROC found by Kim et al. (2025) in noisy external videos should not be placed on one leaderboard as if they estimate the same quantity. The latter is a more demanding and clinically informative test.

Reference standards also vary. Some datasets contain clinician-confirmed diagnoses based on comprehensive evaluation; others contain caregiver-reported labels or labels derived from the same questionnaire family used as predictors. Training against a weak reference standard caps the meaning of model performance. AI can reproduce a label efficiently without showing that the label is correct.

5.2 External validation and distribution shift

External validation asks whether a locked model works in data not used for development. Stronger forms include temporal validation, validation at different sites, geographic and cultural validation, device and language validation, and prospective evaluation in the intended pathway. Each tests a different source of distribution shift.

The evidence consistently shows attenuation as evaluation becomes more realistic. Rajagopalan et al. (2024) declined from internal AUROC 0.895 to 0.790 in independent cohorts. Tariq et al. (2018) reported

a reduction from 0.93 to 0.86 for the newly developed video-feature model. Kim et al. (2025) moved from 0.83 in the principal multisite sample to 0.73 in noisy external recordings. These declines should be expected. A clinical model should be valued for retaining useful performance after such stress, not penalized for revealing it.

External validation is particularly important across cultures. Social gaze, gesture, caregiver interpretation, language, play routines, camera access, and referral pathways are culturally and economically patterned. A model developed in one country should not be presumed valid elsewhere because autism criteria are internationally recognized. Alsharif et al. (2025) usefully foregrounded Arab datasets, but two public datasets and internal cross-validation cannot establish cross-country generalizability or cultural validity.

5.3 Prevalence, calibration, and abstention

Discrimination describes ranking; calibration describes whether predicted probabilities correspond to observed outcomes. A model can have a good AUROC and poor calibration. Yet calibration plots, calibration intercepts and slopes, and decision-curve analyses are rarely reported in the focal autism literature. This omission limits threshold selection and shared decision-making.

Predictive values must be reported at the prevalence of the intended setting. Perochon et al. (2023) achieved AUROC 0.90 but a positive predictive value of 40.6%, meaning that a positive screen still required careful follow-up. This is compatible with a useful screener but incompatible with describing every positive result as a diagnosis. Negative predictive value was high, but safety also depends on who was missed and whether performance is stable in underrepresented subgroups.

Abstention adds a third outcome: “the system cannot determine.” This can protect patients when inputs are poor or presentations are outside the confident region. However, coverage must be reported alongside performance. In Megerian et al. (2022), high determinate-case sensitivity applied to fewer than one third of participants. The appropriate comparison is therefore not simply AI versus clinician accuracy. It is a pathway comparison that asks whether determinate outputs are correct, whether indeterminate cases reach appropriate assessment, and whether total wait time or workload improves.

5.4 Fairness, representativeness, and hidden exclusion

Fairness cannot be established by reporting similar average accuracy for broad demographic categories. Subgroups may be too small to detect meaningful differences, and recorded categories may not capture language, intellectual disability, socioeconomic position, rurality, device quality, gender diversity, or co-occurring conditions. Equal sensitivity can coexist with unequal access to the tool or unequal consequences of error.

Historically diagnosed datasets can also encode ascertainment bias. People whose presentation conforms to established stereotypes are more likely to have been diagnosed and therefore included as positive examples. Girls, racialized children, children with fluent language, and people who mask may be

underdiagnosed or diagnosed later. A model trained on historical labels can learn whom the system has recognized, not the full diversity of autism.

Digital systems introduce an access layer. Home-video tools require a suitable device, connectivity, time, privacy, and confidence following instructions. Voice and text systems require language coverage. Eye-tracking requires hardware and tolerable test procedures. If these requirements are ignored, a nominally scalable tool may concentrate benefits among families already positioned to access care. Han, Weng, and Xu (2026) therefore argue that equity should be assessed across the care pathway rather than treated as a post hoc subgroup statistic.

5.5 Neurodiversity, discourse, and the target of prediction

Technical evaluations often assume that detecting autism earlier is intrinsically beneficial. Earlier identification can facilitate access to communication supports, accommodations, health care, education, and family understanding. Benefit is not automatic, however. It depends on what follows identification and whether services respect autonomy, dignity, and individual goals.

Keyes (2020) showed that autism AI research can reproduce deficit-only and dehumanizing narratives, treating autistic behavior as an object to be detected while excluding autistic accounts of communication. The “double empathy” perspective further suggests that interaction difficulties can be relational rather than located entirely within the autistic person (Erden et al., 2021). Systems trained to measure deviation from neurotypical behavior may pathologize difference, encourage masking, or infer internal states from ambiguous outward behavior.

Participatory governance should therefore begin before model development. Autistic people, including people with differing communication and support needs, should help define intended benefits, unacceptable harms, language, consent processes, outputs, and override mechanisms. Parents and clinicians are essential stakeholders but cannot substitute for autistic participation. A clinically accurate model can still be unacceptable if its purpose, framing, or downstream use undermines autonomy.

5.6 Reproducibility, conflicts of interest, and research integrity

Clinical claims require traceable evidence. Minimum reproducibility includes a clear participant flow, accessible protocol, feature definitions, model version, preprocessing sequence, split logic, missing-data handling, thresholds, complete metrics with confidence intervals, and enough code or executable detail for independent evaluation. Proprietary constraints may limit release of production code, but they do not justify unidentified datasets or irreproducible methods.

The retraction of Jiang’s Scientific Reports article demonstrates the failure mode directly. Editors could not verify the claimed training or validation datasets, underlying code was absent, the framework was inadequately described, and a generated figure contained nonsensical content (Jiang, 2025). The lesson

is broader than one paper: polished diagrams, explainability labels, and high metrics cannot compensate for missing provenance.

Commercial development is not inherently disqualifying, and regulated-device studies may be more clinically rigorous than many academic benchmark studies. Nevertheless, conflicts and sponsor roles should be explicit, and independent replication is especially important when authors or funders have financial interests. Megerian et al. (2022) provided a comparatively strong prospective design but also involved substantial company relationships and a specialized analyst workflow. Decision makers need both facts to assess applicability.

6. A Clinical-Readiness Framework

The evidence supports a five-stage framework for evaluating autism assessment AI. Advancement requires satisfying the earlier stage; a model should not be treated as clinically ready because it performs well at a later technical task while its data or reference standard is weak.

Table 2

Clinical-readiness framework for pediatric autism assessment AI

Stage	Core question	Minimum evidence	Common failure signal
1. Data validity	Do the data and labels represent the intended clinical construct and population?	Independently established reference standard; participant-level splitting; clear provenance; clinically relevant comparator groups; representative recruitment	Questionnaire score or label-derived variable dominates prediction; typical controls only; unclear diagnosis source
2. Model validity	Does the model estimate risk reliably in unseen data from the same intended context?	Locked preprocessing; leakage controls; held-out evaluation; calibration; confidence intervals; missingness and uncertainty handling; reproducible methods	Cross-validation presented as external validation; accuracy alone; unreported thresholds; undocumented preprocessing
3. Clinical validity	Does performance transport to new sites, times, devices, languages, and subgroups?	Geographic and temporal validation; prospective pathway study; PPV/NPV at realistic prevalence; subgroup estimates; explicit coverage	Near-perfect internal performance with no external test; large decline under modest noise; subgroup claims based on very small cells
4. Workflow utility	Does the system improve care rather than merely predict a label?	Comparison with current practice; time, cost, referral, workload, and error outcomes; usability; monitoring and override procedures	No evidence that clinicians act more accurately or efficiently; hidden manual labor; indeterminate cases excluded from headline metrics
5. Equitable benefit	Are benefits accessible and acceptable, and are harms governed?	Participatory design with autistic people; access analysis; consent and privacy safeguards; audit plan; post-deployment surveillance; accountable human decision maker	Deficit-only framing; device or language exclusion; fairness reduced to one demographic table; no mechanism for contesting output

This framework reframes model comparison. An externally validated AUROC of 0.80 with good calibration, transparent abstention, and a demonstrated reduction in wait time may be more useful than 98% internal accuracy on a convenience dataset. Clinical readiness is a property of the full system in context, not of the algorithm alone.

7. Future Directions

7.1 Prospective, multisite pathway trials

Future studies should enroll consecutive children at the actual point of use, lock the model before evaluation, and compare AI-supported care with current practice. Outcomes should include time to assessment, specialist hours, proportion receiving a result, false-negative follow-up, family burden,

costs, and access—not only AUROC. Cluster or stepped-wedge designs may be appropriate when tools are introduced at clinic level.

7.2 Clinically meaningful comparator groups

Evaluation should include the children clinicians find difficult to differentiate: language and developmental delay, attention-deficit/hyperactivity disorder, anxiety, intellectual disability, hearing differences, trauma, and multiple co-occurring conditions. Autism-versus-typical classification is useful for early development but insufficient as the principal claim for diagnostic support.

7.3 Calibration, decision analysis, and transparent uncertainty

Studies should report calibration plots and statistics, precision–recall curves, PPV and NPV at relevant prevalence, decision-curve analysis, and threshold-specific consequences. Indeterminate outputs should be counted in the denominator. Selective-prediction systems should report the accuracy–coverage tradeoff and specify the pathway for every abstained case.

7.4 External validation across culture, language, and technology

Models should be evaluated across countries, languages, racial and ethnic groups, socioeconomic settings, and device conditions. Translation requires more than converting questionnaire text; it requires testing whether tasks, annotations, and behavior interpretations remain valid. Federated and privacy-preserving collaboration may help assemble diverse datasets, but shared governance and harmonized outcome definitions are prerequisites.

7.5 Human-centered explainability

Explainability studies should randomize or compare explanation formats and measure whether they improve error detection, calibrated trust, decision quality, and communication. Explanations should communicate uncertainty and limitations, not merely rank features. Causal language should be prohibited unless supported by an appropriate causal design.

7.6 Participatory and neurodiversity-informed design

Autistic people should participate in priority setting, annotation policy, acceptable-use decisions, interface testing, and governance. Research should include people who use augmentative and alternative communication and people with higher support needs. Outcomes should encompass autonomy, comfort, perceived respect, privacy, and access to useful supports.

7.7 Reproducible and auditable AI

Future research should follow TRIPOD + AI for prediction models and STARD-AI for diagnostic-accuracy studies (Collins et al., 2024; Sounderajah et al., 2025). Protocols and analysis plans should be registered

when feasible.

7.8 Post-deployment surveillance

Performance can drift as referral practices, cameras, demographics, and clinical criteria change. Deployed systems therefore require version control, subgroup monitoring, incident reporting, recalibration rules, and clear responsibility for corrective action. Families and clinicians need a way to challenge or correct outputs. Regulatory authorization should be viewed as the beginning of monitoring, not the end of evaluation.

8. Conclusion

AI can make portions of pediatric autism screening and assessment faster, more structured, and more accessible. The strongest evidence comes from prospective, multisite, and externally tested systems, and these studies generally report more moderate performance than curated-data benchmarks. This pattern is a sign of methodological maturity, not technological failure.

The central conclusion is that accuracy is not clinical readiness. Screening must be distinguished from diagnosis; cross-validation from external validation; feature attribution from causal or clinically meaningful explanation; and technical performance from equitable benefit. Near-term systems should support triage, screening, and clinician decision making while communicating uncertainty and preserving accountable human judgment. Translation requires valid labels, leakage-resistant evaluation, calibration at deployment prevalence, explicit abstention, diverse prospective validation, participatory governance, reproducibility, and evidence that the complete pathway improves outcomes.

References

1. Abbas H, Garberson F, Glover E, Wall DP (2018) Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc 25(8):1000–1007. https://doi.org/10.1093/jamia/ocy039
2. Alsharif NA, Al-Nefaie AH, Ahmad S, Farhah NS (2025) Artificial intelligence-driven diagnosis of autism spectrum disorder in children. Front Med 12:1723320. https://doi.org/10.3389/fmed.2025.1723320
3. Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS (2016) Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. J Child Psychol Psychiatry 57(8):927–937. https://doi.org/10.1111/jcpp.12559
4. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix A-L, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Logullo P (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

5. Erden YJ, Hummerstone H, Rainey S (2021) Automating autism assessment: What AI can bring to the diagnostic process. J Eval Clin Pract 27(3):485–490. https://doi.org/10.1111/jep.13527
6. Han W, Yang X, Li X, Wang J, Liu J, Pang W (2026) Machine learning-based diagnosis of autism spectrum disorder in children and adolescents using eye-tracking data: A systematic review and meta-analysis. Int J Med Informatics 208:106235. https://doi.org/10.1016/j.ijmedinf.2025.106235
7. Han X, Weng L, Xu H (2026) Artificial intelligence, autism care, and health equity: A public health narrative review. Front Public Health 14:1898818. https://doi.org/10.3389/fpubh.2026.1898818
8. Itu II, Islam MS, Malakar NC (2021) *AI for autism—Uses of AI to give a better life to autistic children* [Bachelor's project report, Daffodil International University]. ResearchGate. https://doi.org/10.13140/RG.2.2.10708.17282
9. Jia S-J, Jing J-Q, Yang C-J (2025) A review on autism spectrum disorder screening by artificial intelligence methods. J Autism Dev Disord 55(9):3011–3027. https://doi.org/10.1007/s10803-024-06429-9
10. Jiang S (2025) Retraction note: Bridging the gap: Explainable AI for autism diagnosis and parental support with TabPFNMix and SHAP. Sci Rep 15:43184. https://doi.org/10.1038/s41598-025-31337-y
11. Keyes O (2020) Automating autism: Disability, discourse, and artificial intelligence. J Sociotechnical Critique 1(1):1–31. https://doi.org/10.25779/89bj-j396
12. Kim DY, Do R, Shin Y, Sim H, Kim H, Cho S, Lee G, Park S, Jang B, Lim H, Ha S, Yu J, Choi H, Lee J, Park M-H, Cho A, Yang C-M, Lee D, Yoo H, Kim B-N (2025) Automated AI based identification of autism spectrum disorder from home videos. npj Digit Med 8:607. https://doi.org/10.1038/s41746-025-01993-5
13. Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls CJ, Sohl K, Rouhbakhsh R, Narasimhan A, Romain J, Golla S, Shareef S, Ostrovsky A, Shannon J, Kraft C, Liu-Mayo S, Abbas H, Gal-Szabo DE, Wall DP, Taraman S (2022) Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. npj Digit Med 5:57. https://doi.org/10.1038/s41746-022-00598-6
14. Mertz L (2024) Using AI and ML to predict autism spectrum disorder. IEEE Pulse 15(3):11–15. https://doi.org/10.1109/MPULS.2024.3443489
15. Mohammadi F, Shahrokhi H, Asadzadeh A, Pirmoradi S, Moghtader A, Rezaei-Hachesu P (2025) Artificial intelligence in autism spectrum disorder diagnosis: A scoping review of face, voice, and text analysis methods. Health Sci Rep 8(11):e71476. https://doi.org/10.1002/hsr2.71476
16. Perochon S, Di Martino JM, Carpenter KLH, Compton S, Davis N, Eichner B, Espinosa S, Franz L, Babu K, Sapiro PR, G., Dawson G (2023) Early detection of autism using digital behavioral phenotyping. Nat Med 29(10):2489–2497. https://doi.org/10.1038/s41591-023-02574-3
17. Rajagopalan SS, Zhang Y, Yahia A, Tammimies K (2024) Machine learning prediction of autism spectrum disorder from a minimal set of medical and background information. JAMA Netw Open 7(8):e2429229. https://doi.org/10.1001/jamanetworkopen.2024.29229
18. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D, Bossuyt PM, Darzi A, Ashrafian H (2025) The STARD-AI reporting guideline for

diagnostic accuracy studies using artificial intelligence. Nat Med 31:3283–3289. https://doi.org/10.1038/s41591-025-03953-8

19. Sun C, McEwan A, Boulton KA, Demetriou EA, Sadozai AK, Lampit A, Guastella AJ (2025) Artificial intelligence for tracking social behaviours and supporting an autism spectrum disorder diagnosis: Systematic review and meta-analysis. EBioMedicine 120:105931. https://doi.org/10.1016/j.ebiom.2025.105931
20. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP (2018) Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Med 15(11):e1002705. https://doi.org/10.1371/journal.pmed.1002705