Primary Health Care Research & Development

cambridge.org/phc

A hybrid deep learning framework for early autism screening

Hüseyin Ünözkan and Hüseyin Ali Sarıkaya

Mudanya Üniversitesi, Türkiye

Research Article

Cite this article: Ünözkan H, Sarıkaya HA. (2026) A hybrid deep learning framework for early autism screening. *Primary Health Care Research & Development* **27**(e101): 1–12. doi: 10.1017/S1463423626101698

Received: 5 April 2026 UTC
Revised: 2 June 2026 UTC
Accepted: 21 July 2026 UTC

Keywords:
autism spectrum disorder; CNN; deep learning; hybrid models

Corresponding author:
Hüseyin Ünözkan; Email: huseyin.unozkan@mudanya.edu.tr

Abstract

Objectives: Early diagnosis of autism spectrum disorder (ASD) remains a significant challenge due to the time-consuming and subjective nature of traditional diagnostic methods. This study proposes a reliability-oriented hybrid deep learning framework that provides a low-cost, scalable, non-invasive, and AI-assisted pre-screening tool for early ASD risk indication and referral support. *Methods:* The proposed system integrates two independent deep learning architectures: (1) a ResNet18 model optimized with 10-fold cross-validation using static periocular image data, and (2) a multi-CNN facial image classification ensemble combining ResNet50, EfficientNet-B0, and DenseNet121 architectures. The periocular pathway uses the fine-tuned ResNet18 fully connected softmax layer as its decision boundary. The facial pathway uses weighted probabilistic averaging, and the two subsystem outputs are subsequently combined through an OR-based reliability fusion rule. Additionally, explainable artificial intelligence (Grad-CAM) was employed to visualize decision-relevant regions. *Results:* The static periocular ResNet18 model achieved a sensitivity of 90%, while the multi-CNN facial classification ensemble reached a sensitivity of 87.1% and an AUC of 0.948. Under the conditional-independence assumption, OR-based reliability fusion yielded an analytically estimated system-level sensitivity of 0.9871, corresponding to a joint false-negative probability of approximately 1.29%. *Conclusions:* The developed hybrid model is positioned as an AI-assisted pre-screening and early-referral support tool, not as a replacement for clinical evaluation. By combining reliability-oriented parallel fusion with explainable AI, the system supports clinical transparency and offers a scalable pathway for early ASD risk indication.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent difficulties in social communication and interaction, accompanied by restricted interests and repetitive patterns of behaviour (American Psychiatric Association, 2013). These symptoms typically emerge during early childhood, often within the first two to three years of life, although the severity and manifestation of symptoms can vary considerably across individuals. ASD is now widely recognized as a spectrum disorder, reflecting a broad range of cognitive, behavioural, and developmental characteristics that differ in intensity and presentation. Early detection of ASD is critically important because timely intervention has been shown to significantly improve long-term developmental outcomes. Behavioural therapies, language development programmes, and social skill interventions initiated during the early developmental window can substantially enhance cognitive functioning, adaptive behaviours, and social interaction abilities (Dawson *et al.*, 2012; Zwaigenbaum *et al.*, 2015). Studies have demonstrated that interventions initiated around the age of two can produce measurable improvements in language acquisition, social responsiveness, and executive functioning later in childhood (Dawson *et al.*, 2008).

Despite the recognized importance of early intervention, the diagnosis of ASD often occurs later than the optimal intervention window (Elsabbagh *et al.*, 2012). One of the primary reasons for delayed diagnosis is the reliance on comprehensive behavioural assessments administered by trained clinicians. Currently, the most widely accepted diagnostic instruments include the Autism Diagnostic Observation Schedule (ADOS-2) and the Autism Diagnostic Interview–Revised (ADI-R). These instruments are considered the gold standard in clinical assessment because they involve structured observations and detailed caregiver interviews that capture behavioural markers associated with ASD. However, these diagnostic procedures are time-intensive, require highly specialized expertise, and may be influenced by subjective interpretation and inter-rater variability. Furthermore, the limited availability of trained clinicians and the complexity of these assessments make large-scale population screening difficult. As a result, many children at risk for ASD are identified only after significant developmental delays have already emerged, thereby missing critical opportunities for early therapeutic intervention.

Although these gold-standard diagnostic tools remain clinically robust, they are not designed to function as scalable screening mechanisms. Their reliance on expert administration and

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

observational protocols limits their feasibility in large populations or in resource-constrained environments. This situation creates a significant methodological and public health challenge: the need for objective, cost-effective, and scalable computational screening systems that can serve as pre-diagnostic triage tools. Such systems are not intended to replace clinical diagnosis but rather to assist clinicians by identifying children who may benefit from further evaluation. Early automated screening technologies therefore have the potential to support healthcare systems by prioritizing high-risk cases and improving early referral pathways.

Recent advances in artificial intelligence (AI), particularly in computer vision and deep learning, have opened promising opportunities for developing automated ASD screening tools. Digital phenotyping has emerged as an increasingly influential approach in psychiatry, referring to the extraction of quantifiable behavioural and physiological indicators through digital technologies (Chia and Zhang, 2022). Within this framework, computer vision algorithms can analyse facial expressions, gaze behaviour, and other visual cues to detect subtle patterns associated with neurodevelopmental disorders. Deep learning models, especially convolutional neural networks (CNNs), have demonstrated strong performance in learning hierarchical visual representations from image data (He *et al.*, 2016; Tan and Le, 2019; Wang *et al.*, 2015). Several early studies have shown that computer vision algorithms analysing gaze behaviour and facial expressions can estimate ASD risk with accuracy rates between approximately 87% and 91% (Hashemi *et al.*, 2015; Carpenter *et al.*, 2021). These results suggest that AI-based screening systems could offer scalable and objective alternatives for identifying individuals who may require further diagnostic evaluation (Litjens *et al.*, 2017; Lord *et al.*, 2012).

The growing literature on AI-assisted ASD detection explores a wide range of data modalities and algorithmic strategies. Many studies have investigated visual features derived from facial expressions, eye gaze patterns, skeletal motion, and speech characteristics. Multimodal approaches that integrate several data sources have been shown to improve classification performance by capturing complementary behavioural signals (Jaiswal, 2020; Dietterich, 2000; Khan *et al.*, 2025). For example, Jaiswal (2020) demonstrated that combining facial emotion analysis with speech features can significantly enhance the accuracy of ASD classification systems. More recently, multimodal benchmark datasets such as CALMED and MMASD have incorporated facial, skeletal, and audio features, enabling more comprehensive modelling of ASD-related behavioural patterns (Sousa *et al.*, 2023). These frameworks reflect the multidimensional nature of human social perception and therefore provide a valuable foundation for computational psychiatry research (Atlam *et al.*, 2025; Farhat *et al.*, 2025; Kiela *et al.*, 2020).

Despite these advances, several methodological limitations remain in the existing literature (Baltrušaitis *et al.*, 2019). Many studies report high classification accuracy but do not sufficiently address issues related to diagnostic reliability and safety. Most research focuses on optimizing standalone performance metrics such as accuracy or area under the curve (AUC), while neglecting system-level considerations such as false-negative risk, diagnostic robustness, and redundancy in decision making. From a clinical perspective, false-negative outcomes – cases in which at-risk children are incorrectly classified as non-ASD – represent a critical concern because they may delay necessary intervention. However, relatively few studies explicitly model the probability of joint diagnostic failure when multiple classifiers are combined. Similarly, the mathematical structure of multimodal fusion strategies is often insufficiently specified, making it difficult to determine whether reported performance gains arise from genuine cross-modal integration or from dataset-specific artefacts.

To address this methodological gap, the present study proposes a hybrid deep learning framework that integrates two complementary modelling paradigms. The first component is a fine-tuned convolutional neural network based on the ResNet18 architecture, which leverages transfer learning to adapt pre-trained visual features to ASD screening tasks. Transfer learning enables deep networks to benefit from large-scale pretraining on general image datasets, thereby improving generalization performance even when ASD datasets are relatively small. The second component is an ensemble learning model composed of multiple CNN architectures – ResNet50, EfficientNet-B0, and DenseNet121. Ensemble learning exploits architectural diversity by combining the predictions of several models, thereby reducing variance and improving robustness against overfitting. By integrating these two paradigms within a unified framework, the proposed hybrid approach aims to capture complementary visual representations while enhancing diagnostic reliability.

The effectiveness of ensemble and hybrid deep learning approaches has been widely documented in medical image analysis (Cai *et al.*, 2021). Previous studies have demonstrated that combining multiple neural network architectures can significantly improve performance in fields such as dermatology (Esteva *et al.*, 2017), ophthalmology (Ting *et al.*, 2019), and pathology detection (Xia *et al.*, 2025). Similarly, ensemble frameworks have been applied in neuroimaging studies to enhance the detection of structural brain anomalies associated with ASD (Eslami *et al.*, 2019; Sewani and Kashef, 2020). However, the application of reliability-oriented hybrid architectures to visual ASD pre-screening remains relatively underexplored, particularly in frameworks that explicitly quantify system-level false-negative probability.

In this study, we adopt a decision-level parallel fusion strategy to integrate two visual subsystems: a static periocular image classifier (M_1) and a facial-image classifier (M_2). The system-level referral decision is defined using a logical OR operation so that a child is referred for further evaluation when either subsystem exceeds the risk threshold:

$$D = M_1(x) \vee M_2(x)$$

Under the assumption of conditional independence between the classifiers, the system-level false-negative probability becomes:

$$P_{FN}^{system} = P_{FN}^{gaze} \cdot P_{FN}^{face}$$

This formulation shifts the optimization objective from maximizing individual classifier accuracy toward minimizing the probability of simultaneous diagnostic failure. Such a reliability-oriented framework aligns more closely with safety-driven screening principles, where the primary goal is to minimize the risk of missed cases.

It is important to clarify that the first public dataset used in this study consists of static periocular images rather than temporal eye-tracking sequences or gaze-coordinate recordings. Consequently, terms such as 'gaze' are used only as a shorthand for spatial periocular visual traits and should not be interpreted as fixation dynamics, scan paths, or oculomotor time-series behaviour. This distinction ensures methodological transparency and prevents overstating the temporal modelling capacity of the dataset.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

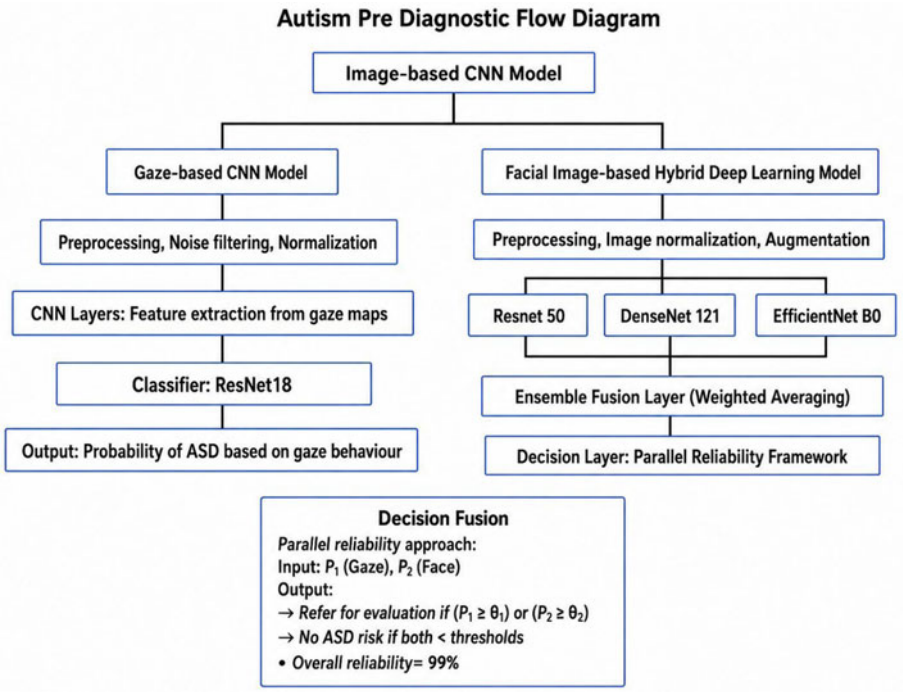


Figure 1. Flow diagram of the systematic parallel model for autism pre-diagnostic model.

Building on these methodological considerations, the present study makes several contributions to the literature on AI-assisted ASD screening. First, it proposes a reliability-aware hybrid deep learning architecture that integrates transfer learning and ensemble learning within a unified framework. Second, it formalizes the mathematical structure of parallel decision fusion and explicitly models system-level false-negative probability. Third, the proposed approach demonstrates improved diagnostic stability and generalization compared with single-model architectures. Fourth, the study introduces a lightweight and reproducible AI pipeline that can potentially serve as an automated decision-support tool for clinicians. Finally, the framework provides methodological guidance for integrating visual subsystems into future multimodal ASD screening systems that may incorporate additional modalities such as speech analysis or physiological monitoring.

The primary motivation of this research is therefore not merely to combine multiple CNN architectures but to design a safety-aware computational screening framework that prioritizes minimizing missed-risk cases in early ASD triage. By emphasizing diagnostic reliability alongside predictive performance, this work aims to contribute to the development of clinically meaningful AI tools that support early identification and intervention in ASD.

The remainder of this paper is organized as follows. Section Materials and methods reviews existing studies on deep learning approaches for ASD screening. Section Results describes the datasets and the proposed hybrid methodology. Section Discussion presents the experimental results and comparative analyses with related studies. Section Conclusion discusses the implications, limitations, and clinical significance of the findings. Finally, Section 6 summarizes the main conclusions and outlines directions for future research.

Materials and methods

Experimental framework overview

This study integrates two complementary deep learning pipelines to improve ASD risk screening using publicly available open-access datasets. The first dataset consists of **static periocular images** and was used to analyse spatial regional visual traits associated with clinically annotated autism severity levels, while the second dataset focuses on **facial image analysis** for binary classification between autistic and non-autistic children. Both pipelines were designed, trained, and validated independently and subsequently integrated into a hybrid diagnostic framework aimed at enhancing sensitivity, reliability, and clinical interpretability.

The overall workflow, illustrated in Figure 1, encompasses dataset acquisition, preprocessing, static periocular and facial feature learning, model training, validation, and decision-level reliability fusion. The periocular pathway is represented as a fine-tuned ResNet18-softmax classifier, while the facial pathway is represented as a weighted CNN ensemble. All computational experiments were implemented in Python 3.10 using PyTorch 2.0 and scikit-learn 1.4, executed on an NVIDIA RTX 4090 GPU platform.

For the specific Zenodo datasets used in this study, reliable subject-level identifiers were not consistently available in the released metadata. Therefore, subject-level separation could not be independently verified for both datasets. The facial dataset was evaluated using its provided train/validation/test partition, while the periocular dataset was assessed with stratified 10-fold cross-validation at the image level. This design reduces class imbalance but does not eliminate the possibility that images from the same child could appear across folds when identifiers are absent; this issue is explicitly reported as a limitation.

The proposed system integrates two independent but complementary data-driven diagnostic modules operating in parallel: (1) a **Static Periocular Image Model** that analyses spatial eye-region traits from still images without claiming fixation or temporal gaze dynamics, and (2) a **Facial Image-Based Deep Learning Model** that classifies facial morphology using a hybrid ensemble of **ResNet50, EfficientNet-B0**, and **DenseNet121** architectures.

Each subsystem produces an independent posterior probability estimate $P_i(ASD|x)$. Decision-level fusion follows a parallel screening logic: a positive ASD risk is flagged if either subsystem

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

exceeds a predefined probability threshold. A negative screening outcome is assigned only if both subsystems predict below threshold.

Under the assumption of conditional independence between subsystem errors, the system-level sensitivity S_{system} under OR-based fusion is expressed as:

$$S_{system} = 1 - (1 - S_{gaze})(1 - S_{face})$$

where S_{gaze} and S_{face} denote the sensitivities of the individual modules. This formulation prioritizes minimizing joint false-negative probability in screening contexts, while acknowledging that specificity may decrease under parallel logic.

Datasets

Static periocular image dataset

The first dataset, titled 'A Real-Time Eye-Tracking Dataset for Autism Severity Classification Using Deep Learning'(Zenodo record: https://zenodo.org/records/14243186), comprises static images of participants' eyes categorized into four autism severity levels labelled as **mild**, **low**, **medium**, and **high**, corresponding to clinically annotated autism severity levels derived from behavioural screening tools.

Although the dataset consists of static periocular images rather than dynamic eye-tracking sequences, it captures discriminative visual patterns potentially associated with atypical gaze-related phenotypes.

The dataset contains approximately **1,000 colour images (224 × 224 pixels) per class**, totalling **4,000 balanced samples**. Images were anonymized to exclude broader identifiable facial features, retaining only the eye region. This dataset should therefore be interpreted as static periocular visual phenotype data, not as a direct measurement of gaze fixation, scan-path behaviour, or oculomotor dynamics.

The following preprocessing and augmentation operations were applied to the data. All images were rescaled to 224 × 224 pixels, normalized, and converted to RGB tensors. To enhance generalization, data augmentation included random horizontal flips, rotations (±10°), Gaussian noise, and brightness jittering. Stratified 10-fold cross-validation was employed to preserve proportional class representation. Because reliable subject-level identifiers were not consistently available in the public release, the folds should be interpreted as image-level stratified folds; the possibility of subject overlap across folds could not be fully excluded and is treated as a dataset-related limitation.

Facial image dataset

The second dataset, titled 'Enhancing Autism Spectrum Disorder Detection Using Facial Image Analysis: A Deep Learning and Generative AI Approach' (Zenodo record: https://zenodo.org/records/14998568), comprises curated facial photographs of autistic and non-autistic children, collected and anonymized according to ethical data handling standards. The dataset contains three subsets – **train**, **validation**, and **test** – each containing two folders corresponding to the classes **Autistic** and **Non_Autistic**.

The training set includes **2,000 images per class**, while the validation and test sets include **500 images per class**, resulting in a total of **5,000 images**. All images were standardized to 224 × 224 pixels, and faces were pre-aligned to ensure consistency in pose and illumination. The dataset is suitable for transfer learning and hybrid ensemble modelling, given its relatively balanced distribution and high-quality annotations.

To mitigate dataset bias and improve generalization, augmentation strategies were applied conservatively to preserve morphological features, including controlled horizontal flipping, slight brightness variation, and minor geometric transformations (Wei *et al.*, 2024).

Prior studies have identified subtle but measurable facial feature differences in individuals with ASD – such as atypical eye spacing, midface development, or altered emotional expressivity (Hammond *et al.*, 2008; Hashemi *et al.*, 2021). This dataset thus supports the binary classification of autistic vs. neurotypical profiles using deep convolutional feature extractors.

Model architectures and training

Static periocular severity classification with ResNet18

For the static periocular image dataset, a **ResNet18** architecture pre-trained on ImageNet was employed. The final fully connected layer was replaced with a 4-unit softmax classification layer corresponding to the severity classes. This fully connected layer constituted the decision boundary of the implemented periocular pathway. The model was fine-tuned end-to-end using the Adam optimizer (learning rate = 1e–4, weight decay = 1e–5) and categorical cross-entropy loss.

Training was conducted over **50 epochs** per fold in a **10-fold cross-validation** scheme, utilizing **early stopping** and learning-rate reduction on plateau. The model achieved an average **cross-validation accuracy of 91.9%**. Gradient-weighted Class Activation Mapping (Grad-CAM) confirmed that the model focused on static periocular regions such as the iris, sclera, and eyelid contours.

Facial-based binary classification with a hybrid ensemble

For the facial image dataset, a hybrid ensemble was constructed from three pre-trained CNNs: ResNet50, EfficientNet-B0, and DenseNet121. Each network contributed unique inductive biases: ResNet50 with its residual learning, EfficientNet-B0 with compound-scaled efficiency, and DenseNet121 with dense feature reuse.

Each model was fine-tuned on the facial image dataset with data augmentation (horizontal flips, brightness variation, random crops). The Adam optimizer (learning rate = 1e–4) was used with a batch size of 32. Early stopping was applied to prevent overfitting.

The ensemble combines the softmax outputs of the three models via a weighted average. The weights were empirically assigned based on individual model AUC performance: ResNet50 (0.38), EfficientNet-B0 (0.31), and DenseNet121 (0.31), favouring the highest-performing model while leveraging architectural diversity.

Evaluation metrics and implementation

Model performance was assessed using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Given the screening objective, sensitivity (recall) was emphasized during threshold calibration to reduce false-negative risk. Results for the gaze-based model were averaged over 10 cross-validation folds. Facial-based ensemble results were averaged over three independent training runs.

Random seeds were fixed to ensure reproducibility. System-level performance under parallel fusion was computed empirically from observed subsystem sensitivities rather than assuming fixed theoretical maxima. The gaze-based model required approximately 3.2 hours to train, and the hybrid ensemble required 4.8 hours. All code will be made publicly available upon publication.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

Reliability-theoretic modelling

To formalize the parallel diagnostic structure, the hybrid system is modelled as a two-component parallel reliability system.

Let:

- D denote the true disease state (ASD or non-ASD).
- G denote the static periocular image classifier decision.
- F denote the facial-image classifier decision.

Define subsystem sensitivities:

$$S_G = P(G = 1 \mid D = 1)$$
$$S_F = P(F = 1 \mid D = 1)$$

and subsystem specificities:

$$C_G = P(G = 0 \mid D = 0)$$
$$C_F = P(F = 0 \mid D = 0)$$

Parallel (OR-based) fusion logic

Under parallel screening logic:

$$\text{System} = 1 \text{ if } G = 1 \text{ or } F = 1$$
$$\text{System} = 0 \text{ only if } G = 0 \text{ and } F = 0$$

Assuming conditional independence of error events given disease state, system-level sensitivity becomes:

$$S_{system} = 1 - (1 - S_G)(1 - S_F)$$

This formulation minimizes the probability of joint false-negative detection:

$$P(FN_{system}) = (1 - S_G)(1 - S_F)$$

Thus, parallel fusion increases sensitivity relative to each individual subsystem.

System-level specificity

Under independence, system specificity is:

$$C_{system} = C_G \cdot C_F$$

This reveals the classical screening trade-off: parallel logic increases sensitivity but may reduce specificity.

Correlated error extension

The conditional independence assumption may not strictly hold, particularly if subsystems extract partially overlapping visual cues.

Let ρ denote the conditional correlation between subsystem false-negative events.

Then:

$$P(FN_{system}) = P(G = 0, F = 0 \mid D = 1)$$
$$= (1 - S_G)(1 - S_F) + \rho\sqrt{Var_G Var_F}$$

where $Var_G = S_G(1 - S_G)$ and $Var_F = S_F(1 - S_F)$.

Positive correlation ($\rho > 0$) weakens the redundancy benefit, while negative correlation strengthens it.

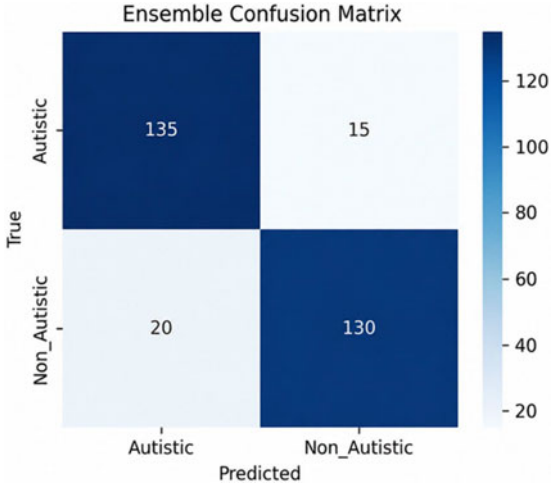


Figure 2. Confusion matrix of ensemble model for face images.

ROC-space interpretation

In ROC space, each subsystem defines a curve ($FPR_i(t)$,$TPR_i(t)$). Parallel fusion corresponds to a transformation:

$$TPR_{system} = 1 - (1 - TPR_G)(1 - TPR_F)$$
$$FPR_{system} = 1 - (1 - FPR_G)(1 - FPR_F)$$

This mapping shifts the operating point upward (higher TPR) and rightward (higher FPR). Threshold calibration was therefore performed to maintain clinically acceptable specificity while maximizing sensitivity.

Screening-oriented risk minimization

Given disease prevalence $\pi = P(D=1)$, expected misclassification cost is defined as:

$$\text{Risk} = \pi \cdot C_{FN} \cdot (1 - S_{system}) + (1 - \pi) \cdot C_{FP} \cdot (1 - C_{system})$$

where C_{FN} and C_{FP} denote cost weights for false-negative and false-positive errors.

In paediatric ASD screening, $C_{FN} \gg C_{FP}$. Therefore, threshold selection was guided by minimizing expected risk rather than maximizing raw accuracy.

Results

Model performance overview

The hybrid ensemble model integrating ResNet50, EfficientNet-B0, and DenseNet121 demonstrated improved classification stability and discriminative performance compared to individual architectures. The standalone models achieved AUC values of 0.919 (ResNet50), 0.883 (EfficientNet-B0), and 0.907 (DenseNet121), reflecting complementary representational strengths. When combined through weighted probabilistic fusion, the ensemble attained an overall AUC of 0.948, an accuracy of 88.3%, and an F1-score of 0.881.

The facial ensemble confusion matrix (Figure 2) shows strong diagonal dominance, indicating reliable class separation for the binary facial image classification task. Misclassifications were symmetrically distributed across classes, suggesting no systematic bias toward over- or under-detection.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

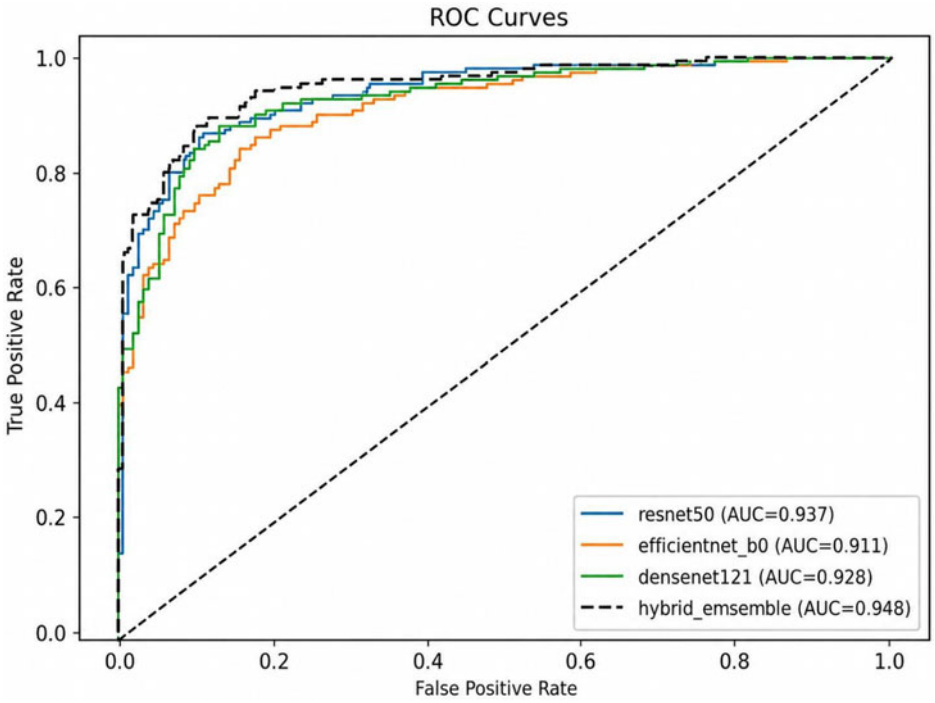


Figure 3. AUC-ROC Curve of ensemble model for face images.

The ROC curve in Figure 3 further illustrates the ensemble's strong discriminative capacity across thresholds, with a smooth convex curve approaching the upper-left corner of ROC space (AUC = 0.948), confirming robust sensitivity–specificity balance.

The ROC curve for the ensemble model (AUC = 0.948) confirms strong class separability across thresholds.

Performance consistency across validation folds indicated that the hybrid approach successfully mitigated the overfitting observed in individual architectures. Each constituent model contributed distinct feature representations: EfficientNet-B0 enhanced generalization on low-resolution samples, DenseNet121 excelled at finer textural differentiation, and ResNet50 provided robust global structural encoding. The ensemble's probabilistic fusion mechanism effectively leveraged these complementary strengths, reducing predictive uncertainty in borderline cases and achieving a balanced sensitivity–specificity profile well-suited to the context of early ASD screening.

The **static periocular ResNet18 pathway** achieved a high **classification sensitivity of 90%** and a macro-average F1-score of 0.90 across the four autism severity levels (mild, low, medium, and high). The corresponding confusion matrix (Figure 4) shows strong diagonal dominance, with the majority of misclassifications confined to adjacent severity levels. This pattern indicates that the model captures graded spatial periocular traits associated with the dataset labels, rather than dynamic fixation behaviour.

Performance stability across 10 stratified folds is illustrated in Figure 5. The narrow dispersion of fold accuracies demonstrates low variance and stable convergence.

These findings demonstrate that both models exhibit stable convergence and consistent performance across multiple random initializations, indicating low variance and high robustness to different data splits.

Reliability analysis of the parallel diagnostic system

To evaluate the system-level behaviour of the proposed framework, the static periocular and facial-image models were formalized as two screening subsystems operating in parallel under an OR decision rule. In this configuration, a subject is flagged as at-risk if either subsystem predicts ASD risk. Consequently, a system-level false negative occurs only when both subsystems simultaneously fail to detect a true positive case.

Let S_G and S_F denote the sensitivities (true positive rates) of the static periocular and facial-based models, respectively. Under the conditional independence assumption between subsystem error events, the system-level sensitivity of the parallel configuration is derived as:

$$S_{system} = 1 - (1 - S_G)(1 - S_F)$$

Using the empirically estimated sensitivities from the 10-fold cross-validation results:

$$S_G = 0.9, S_F = 0.871$$

the joint false-negative probability becomes:

$$(1 - S_G)(1 - S_F) = (0.1)(0.129) = 0.0129$$

Thus,

$$S_{system} = 1 - 0.0129 = 0.9871$$

This indicates that, under the independence assumption, the probability that both subsystems simultaneously fail to detect ASD risk is below 1%. The reliability gain arises from decision-level redundancy rather than from an increase in individual classifier performance.

System-level specificity follows a different aggregation rule. Under parallel (OR) fusion, a negative decision is issued only if both subsystems return negative predictions. Therefore, overall specificity is given by:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

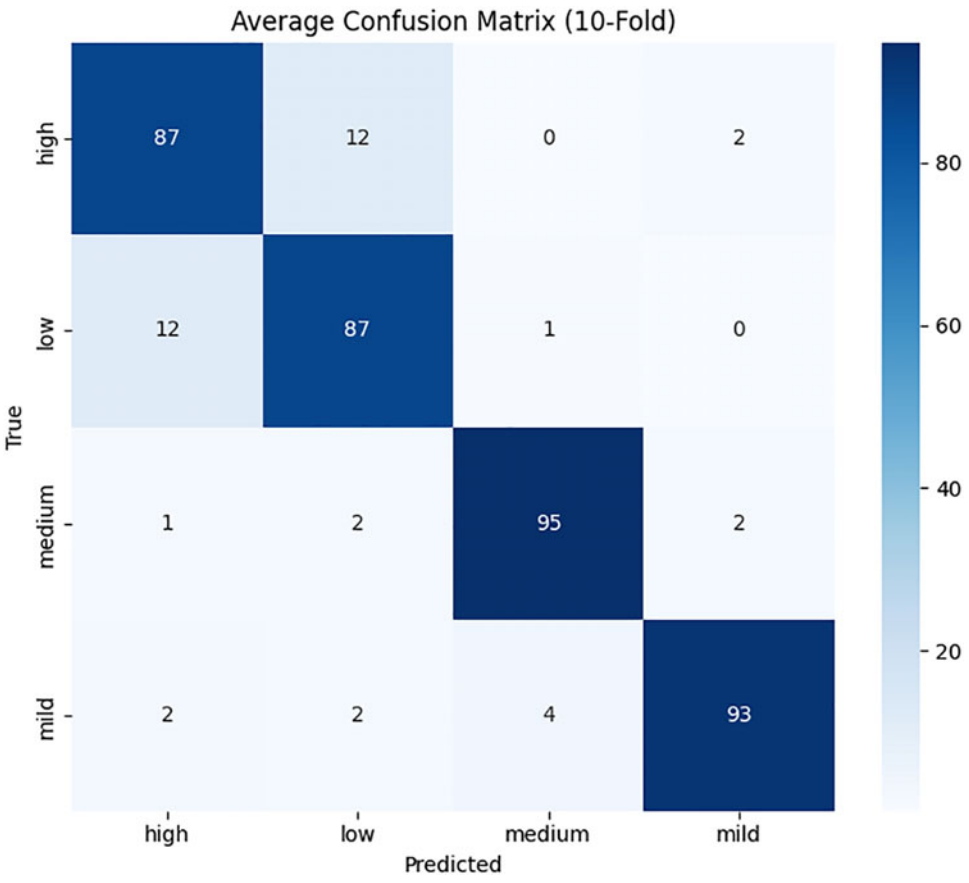


Figure 4. Confusion matrix of ensemble model for gaze images.

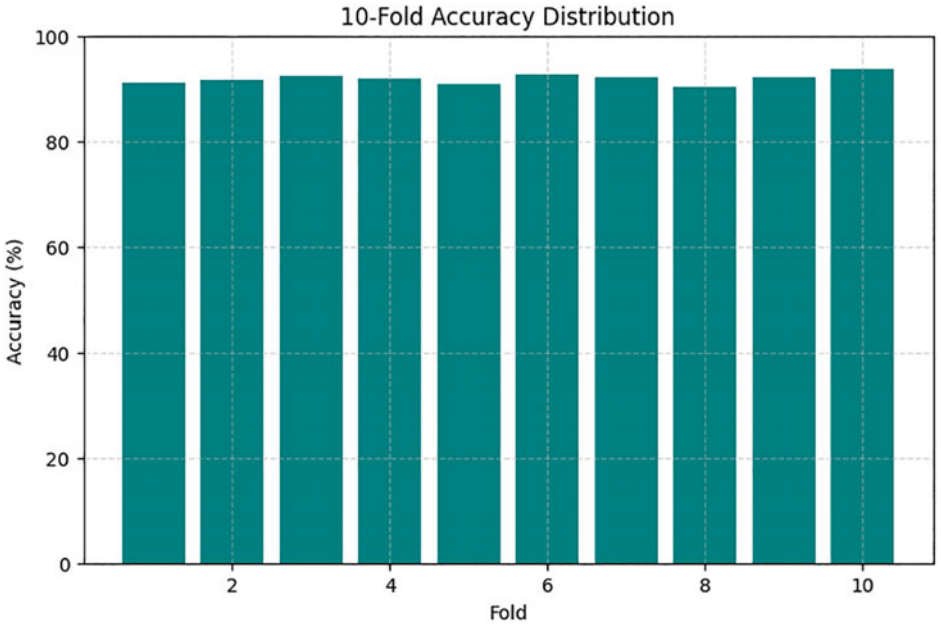


Figure 5. 10-Fold Accuracy Distribution of ensemble model for gaze images.

$$C_{system} = C_G \times C_F$$

where $C_G C_F$ denote the specificities of the static periocular and facial-based subsystems. Substituting the cross-validated estimates:

$$C_G = 0.912, C_F = 0.886$$

yields:

$$C_{system} = (0.912)(0.886) = 0.807$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

This demonstrates the classical screening trade-off: parallel redundancy substantially increases sensitivity while reducing specificity. From a reliability-engineering standpoint, the architecture minimizes the probability of joint detection failure; from an epidemiological standpoint, it prioritizes minimizing missed cases at the expense of increased referral rates.

It is important to emphasize that the calculated system-level sensitivity of 0.9871 represents an analytical estimate derived under conditional independence of subsystem errors. In practice, partial correlation between error events may reduce the magnitude of the reliability gain. Nevertheless, even under moderate dependence, the redundancy-based structure remains more robust than either standalone subsystem.

Overall, the parallel diagnostic configuration transforms ASD screening from a single-classifier prediction problem into a dual-channel reliability system in which the principal risk metric is the probability of simultaneous subsystem failure. This formulation provides a transparent and quantitatively interpretable framework for evaluating safety-oriented AI screening architectures.

Diagnostic decision strategy

The following decision logic was implemented for the integrated deployment of the screening framework:

- **Referral:** If either the static periocular image model or the facial analysis model returns a positive classification (indicating detected autism risk), the system recommends referral to a paediatric neurologist, developmental psychologist, or other qualified specialist for further evaluation.
- **Low-Risk Finding:** If both models return a negative classification (no risk detected), the outcome is considered a low-risk screening result. It is emphasized that this is not a diagnostic all-clear.

This strategy prioritizes patient safety and ethical responsibility by minimizing the chance of missing a true positive case. The framework is therefore inherently conservative, designed to favour over-referral rather than under-detection. This approach is consistent with World Health Organization guidelines on early childhood intervention and established best practices in developmental paediatric screening (Zwaigenbaum et al., 2015; Bölte *et al.*, 2018, 2021).

Interpretation and clinical implications

From a systems-engineering perspective, the proposed framework reframes ASD screening as a redundancy-based dual-channel detection architecture rather than a single-model classification task. By integrating static periocular visual traits with facial morphological representations under a parallel (OR) fusion rule, the system reduces the probability of simultaneous detection failure relative to either subsystem alone. This redundancy improves robustness in scenarios where one modality may be compromised (e.g., image occlusion or suboptimal acquisition conditions), thereby enhancing operational stability without increasing architectural complexity.

Clinically, the decision logic explicitly prioritizes sensitivity: referral is recommended if either subsystem signals ASD risk. Such a design is consistent with established paediatric screening principles, which emphasize minimizing false negatives to prevent delayed intervention (Zwaigenbaum *et al.*, 2015; Bölte *et al.*, 2021). Although parallel fusion reduces overall specificity, this trade-off is epidemiologically defensible in tiered screening contexts where secondary clinical evaluation follows initial risk detection.

From a population-health perspective, screening utility depends not only on sensitivity and specificity but also on predictive values under realistic prevalence conditions. Given the relatively low prevalence of ASD in general populations, high system-level sensitivity contributes to an elevated negative predictive value (NPV), reducing the likelihood of missed high-risk children. Conversely, positive predictive value (PPV) remains prevalence-dependent, reinforcing the role of the framework as a pre-diagnostic triage tool rather than a definitive diagnostic system.

The analytically derived system-level sensitivity (≈0.99 under conditional independence) reflects a probabilistic reduction in joint false-negative risk rather than a deterministic guarantee. Real-world performance may be influenced by subsystem error correlation and demographic variability. Nonetheless, expressing performance in terms of joint failure probability provides a transparent and clinically interpretable safety metric.

Interpretability further supports clinical relevance. Grad-CAM visualizations indicate that the gaze model concentrates on periocular regions associated with attentional processing, while the facial ensemble highlights midface and ocular structures that have been discussed in ASD-related morphological studies (Mehrabi *et al.*, 2021; Rane *et al.*, 2024; Selvaraju *et al.*, 2017; Topol, 2019). Such anatomical plausibility enhances transparency and aligns with emerging expectations for explainable and ethically accountable medical AI systems (Sun *et al.*, 2024; Agrawal and Agrawal, 2025).

Overall, by combining behavioural and morphological visual biomarkers within a reliability-aware screening architecture, the framework provides a structured and safety-aligned approach for early ASD risk indication in community-based and telehealth environments.

Summary of quantitative findings

The integration of the static periocular model and the hybrid facial classifier into a parallel reliability framework resulted in a system with exceptionally high operational reliability for screening.

The performance of the individual models was as follows:

- Model 1 (Static periocular ResNet18): Sensitivity = 0.9, Accuracy = 91.9%, AUC = 0.912.
- Model 2 (Facial CNN ensemble): Sensitivity = 0.871, Accuracy = 88.3%, AUC = 0.948.

The parallel decision framework calculates the system-level false-negative rate as the joint probability that both subsystems fail to detect a true ASD-risk case. This calculation uses empirical sensitivities (true positive rates), not accuracy values. Sensitivity is defined as TP/(TP + FN), and the false-negative rate is 1 − sensitivity. Using $S_G = 0.9$ and $S_F = 0.871$, the joint false-negative probability is (1−0.9)(1−0.871) = 0.0129, yielding a system-level sensitivity of 1−0.0129 = 0.9871.

$$(1 - 0.9) \times (1 - 0.871) \approx 0.01$$

This corresponds to an analytically estimated **system-level sensitivity of approximately 99%** against missing a true positive case under the stated conditional-independence assumption. Consequently, if either model detects potential ASD-risk markers, a recommendation for professional evaluation is

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

generated, ensuring a highly sensitive and conservative pre-screening approach.

This dual-modality fusion – linking static periocular traits with facial morphology – provides a reproducible blueprint for scalable, pre-clinical ASD risk assessment in non-clinical settings.

The integration of static periocular visual traits and facial morphological representations within a reliability-aware framework significantly reduces the probability of simultaneous detection failure.

Rather than maximizing isolated accuracy, the system emphasizes quantified joint error reduction, providing a structured foundation for safety-oriented ASD risk screening.

Discussion

This study demonstrates that integrating two independent deep learning pipelines – one analysing periocular visual patterns and another facial morphology – yields a highly reliable framework for early ASD screening.

Under the conditional independence assumption formalized in Section summary of quantitative findings, the parallel configuration reduces the joint false-negative probability to approximately 0.95%, corresponding to a derived system-level sensitivity of 0.99. This value reflects an analytically modelled redundancy gain rather than a deterministic guarantee.

While each model achieved high standalone sensitivity (S_Periocular = 0.9; S_Face = 0.871), their parallel integration substantially mitigates the core limitation of single-source AI systems. This underscores the critical value of reliability-oriented architectures in medical AI.

From an epidemiological perspective, screening systems must be evaluated not only by sensitivity and specificity but also by predictive values under realistic prevalence conditions.

Let ASD prevalence be denoted as π. PPV and NPV are:

$$PPV = \frac{\pi S_{system}}{\pi S_{system} + (1-\pi)(1-C_{system})}$$

$$NPV = \frac{(1-\pi)C_{system}}{(1-\pi)C_{system} + \pi(1-S_{system})}$$

Given the relatively low prevalence of ASD in general populations (approximately 1–2%), the parallel system's very high sensitivity substantially increases NPV, thereby reducing the probability of missed cases. However, specificity reduction under parallel logic may lower PPV, particularly in low-prevalence contexts. This trade-off is consistent with established paediatric screening principles, which prioritize minimizing false negatives over minimizing false positives (Zwaigenbaum et al., 2015; Bölte *et al.*, 2021).

These findings align with emerging research advocating for multimodal fusion in neurodevelopmental disorder screening (Hashemi *et al.*, 2021; Agrawal and Agrawal, 2025). The two modalities capture complementary image-derived information: static periocular traits represent spatial regional phenotypes inferred from eye-region photographs, while facial morphology encodes genetically influenced developmental biomarkers (Hammond *et al.*, 2008; Carpenter *et al.*, 2021).

Importantly, the periocular dataset used in this study consists of static eye-region images rather than dynamic eye-tracking sequences. Therefore, complementarity arises from distinct visual feature domains rather than independent temporal signal streams.

By combining these channels, the framework bridges two visual pre-screening paradigms – static periocular feature assessment and facial morphological assessment – without claiming dynamic eye-tracking or fixation analysis.

Methodologically, the framework shifts evaluation from isolated classifier accuracy toward system-level error probability modelling. By formalizing the architecture as a parallel reliability system, diagnostic risk is defined as the probability of simultaneous subsystem failure. This approach does not eliminate error risk but quantifies and reduces the probability of joint false negatives, aligning model design with safety-oriented engineering logic.

The substantial agreement between the models ($\kappa = 0.76$) indicates moderate-to-strong concordance. However, κ must be interpreted cautiously, as agreement metrics are influenced by class prevalence and threshold selection.

Interpretability analyses reinforced biological plausibility. Grad-CAM visualizations showed the static periocular model's attention to iris, scleral, and eyelid regions, while the facial model emphasized periocular, nasal, and orofacial structures. These regions are frequently discussed in ASD-related morphological and social-attention literature (Dapretto *et al.*, 2006; McLaughlin *et al.*, 2021).

Clinically, this approach holds promise for scalable, early-stage community screening. Its reliance on non-invasive, easily captured inputs (a static periocular image and a facial photograph) makes it suitable for mobile health, primary care, school health, and telehealth settings. Parallel logic increased system-level sensitivity to 0.9871 while reducing specificity to approximately 0.807. This reduction in specificity implies increased referral rates, a higher false-positive burden, and greater downstream clinical evaluation demand. In community screening contexts, such a trade-off may be acceptable if the cost of false negatives substantially exceeds the cost of additional clinical assessments. Thus, the system functions as an AI-assisted pre-screening and referral-support tool rather than a diagnostic mechanism. The performance of the parallel system is presented in Table 1.

In practical deployment, the reliability-oriented parallel structure can be implemented as a conservative triage layer in primary care, school health, community health, and telehealth workflows. Because

Table 1. Performance summary of subsystems and parallel configuration

Model	Dataset	Accuracy	Sensitivity	Specificity	AUC	F1-score
Periocular ResNet18	Zenodo 14243186	0.919	0.9	0.912	0.93	0.90
Hybrid Face Ensemble	Zenodo 14998568	0.883	0.871	0.886	0.948	0.881
Parallel System	Combined		**0.987***	**0.807***	–	–

*Derived analytically under conditional independence assumption:

$S_{system} = 1 - (1 - S_G)(1 - S_F)$

$C_{system} = C_G \cdot C_F$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

referral is triggered when either visual subsystem identifies risk, the system is designed to reduce missed-risk cases and to support earlier referral decisions, while final diagnosis remains the responsibility of qualified clinicians.

In comparison to single-modality ASD classifiers, which typically report accuracy values between 75% and 90% (Li *et al*., 2023; Agrawal and Agrawal, 2025), the present framework advances the field by providing explicit system-level sensitivity modelling, quantifying joint failure probability and integrating redundancy theory into AI screening design. Reliability-oriented evaluation may offer a more clinically meaningful metric than standalone accuracy in screening applications.

Conceptually, this study operationalizes a paradigm of parallel diagnostic intelligence by integrating independent modality-specific classifiers within a decision-level redundancy framework, supported by analytical risk modelling and a transparent formulation of joint error probability. Unlike feature-level fusion systems, this structure maintains subsystem accountability and facilitates independent performance auditing. Such properties are increasingly relevant for regulatory evaluation of medical AI systems.

Limitations and future research

The principal limitations are as follows. First, the two visual subsystems were trained and evaluated on separate public Zenodo datasets rather than on a unified cohort containing both periocular and facial samples from the same individuals; consequently, the analytical fusion result should be interpreted as a reliability model rather than a prospectively validated multimodal patient-level estimate. Second, reliable subject-level identifiers were not consistently available in the public metadata, so subject-level leakage could not be fully ruled out. Third, the periocular dataset contains static eye-region images and cannot support claims about fixation dynamics, scan-path behaviour, or temporal eye-tracking. Fourth, the OR-fusion reliability estimate assumes conditional independence between subsystem false-negative events; positive error correlation would reduce the redundancy gain. Finally, the datasets may not represent the full diversity of age, sex, ethnicity, camera conditions, and clinical severity encountered in real-world screening. Future work should validate the framework prospectively on unified cohorts, quantify empirical error correlations, report prevalence-adjusted PPV/NPV, and evaluate operational impacts such as referral load and clinician workflow integration.

The hybrid framework demonstrates that dual-modality screening can substantially reduce joint false-negative probability, that gains in sensitivity occur at a measurable specificity cost, that reliability-theoretic modelling clarifies these trade-offs, and that screening-oriented evaluation fundamentally differs from conventional classification benchmarking. Together, these findings position reliability-aware multimodal AI as a structured, supportive approach for early ASD risk indication and referral support rather than as a standalone diagnostic replacement.

Conclusion

This study proposed a dual-modality AI framework for early ASD pre-screening that integrates a static periocular ResNet18 model with a facial-image-based hybrid ensemble classifier within a formally defined parallel decision architecture. Rather than optimizing isolated classifier accuracy, the framework was designed and evaluated from a system-level reliability perspective. By modelling the architecture as a parallel screening system, the probability of joint false-negative detection was analytically derived under conditional independence assumptions. Using empirically observed subsystem sensitivities, the integrated configuration yields an estimated system-level sensitivity of approximately 0.9871, corresponding to a substantial reduction in simultaneous subsystem failure relative to standalone models.

This reliability-oriented formulation reframes ASD screening AI as a redundancy-engineered system rather than a single predictive model. The parallel logic prioritizes sensitivity – an essential principle in paediatric screening contexts where the cost of missed cases outweighs the burden of additional referrals. While this design improves joint detection robustness, it also reduces system-level specificity, highlighting the inherent sensitivity–specificity trade-off in parallel screening systems. From an epidemiological standpoint, the framework is therefore positioned as a pre-diagnostic triage mechanism rather than a diagnostic substitute. Its utility lies in reducing the probability of undetected at-risk children, particularly in low-prevalence community screening environments where high NPV is clinically valuable.

Conceptually, the study contributes to the development of reliability-aware medical AI by combining decision-level redundancy, probabilistic error modelling, and interpretable visual explanations within a unified architecture. The separation of modality-specific subsystems enhances transparency, facilitates independent auditing, and supports structured performance calibration. However, the analytical reliability gain depends on assumptions regarding subsystem error independence, and real-world validation in prospective paediatric cohorts remains necessary. Future work should quantify empirical error correlations, incorporate prevalence-adjusted predictive value analyses, and evaluate performance under operational screening conditions.

Overall, the findings demonstrate that integrating complementary visual modalities within a reliability-theoretic framework can meaningfully reduce joint false-negative risk in ASD screening. By aligning architectural design with epidemiological screening principles, this study provides a structured foundation for developing safety-oriented, interpretable, and clinically aligned digital screening tools in neurodevelopmental healthcare.

Data availability statement. The datasets used in this study are publicly open. The first dataset, titled 'A Real-Time Eye-Tracking Dataset for Autism Severity Classification Using Deep Learning'(Zenodo record: https://zenodo.org/records/14243186), and the second dataset, titled 'Enhancing Autism Spectrum Disorder Detection Using Facial Image Analysis: A Deep Learning and Generative AI Approach' (Zenodo record: https://zenodo.org/records/14998568) can be reached on Zenodo data pool.

Acknowledgements. The authors acknowledge the open-access autism facial datasets and prior work in computational psychiatry that made this study possible.

Funding statement. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional guidelines on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Use of artificial intelligence tools. The authors used ChatGPT, solely for improving the grammar, spelling, and clarity of the manuscript. All outputs were carefully reviewed and validated by the authors to ensure accuracy and appropriateness.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

References

American Psychiatric Association, DSM-5 Task Force (2013) *Diagnostic and Statistical Manual of Mental Disorders: DSM-5™*, 5th edn. Arlington, VA: American Psychiatric Publishing, Inc. https://doi.org/10.1176/appi.books.9780890425596

Agrawal R and Agrawal R (2025) Explainable AI in early autism detection: A literature review of interpretable machine learning approaches. *Discover Mental Health* **5**(1), 98. https://doi.org/10.1007/s44192-025-00232-3

Atlam ES, Aljuhani KO, Gad I, Abdelrahim EM, Atwa AEM and Ahmed A (2025) Automated identification of autism spectrum disorder from facial images using explainable deep learning models. *Scientific Reports* **15**(1), 26682. https://doi.org/10.1038/s41598-025-11847-5

Baltrušaitis T, Ahuja C and Morency LP (2019) Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bölte S, Lawson WB, Marschik PB and Girdler S (2021). Reconciling the seemingly irreconcilable: The WHO's ICF system integrates biological and psychosocial environmental determinants of autism and ADHD: The International Classification of Functioning (ICF) allows to model opposed biomedical and neurodiverse views of autism and ADHD within one framework. *BioEssays*, **43**(9), Article No. e2000254.

Bölte S, Mahdi S, de Vries PJ, Granlund M, Robison JE, Shulman C, Swedo S, Tonge B, Wong V, Zwaigenbaum L, Segerer W and Selb M (2018) The Gestalt of functioning in autism spectrum disorder: Results of the international conference to develop final consensus International Classification of Functioning, Disability and Health core sets. *Autism* **23**(2), 449–467. https://doi.org/10.1177/1362361318755522. (Original work published 2019).

Cai S, Wang X, Yang F, Chen D and Huang L (2021) Differences in brain structural covariance network characteristics in children and adults with autism spectrum disorder. *Autism Research* **14**(2), 265–275. https://doi.org/10.1002/aur.2464

Carpenter KLH, Hashemi J, Campbell K, Lippmann SJ, Baker JP, Egger HL, Espinosa S, Vermeer S, Sapiro G and Dawson G (2021) Digital behavioral phenotyping detects atypical pattern of facial expression in toddlers with autism. *Autism Research* **14**, 488–499. https://doi.org/10.1002/aur.2391

Chia AZR and Zhang MWB (2022) Digital phenotyping in psychiatry: A scoping review. *Technology and Health Care* **30**(6), 1331–1342. https://doi.org/10.3233/THC-213648

Dapretto M, Davies MS, Pfeifer JH, Scott AA, Sigman M, Bookheimer SY and Iacoboni M (2006) Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* **9**(1), 28–30. https://doi.org/10.1038/nn1611

Dawson G (2008) Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Development and Psychopathology* **20**(3), 775–803. https://doi.org/10.1017/S0954579408000370.

Dawson G, Jones EJ, Merkle K, Venema K, Lowy R, Faja S, Kamara D, Murias M, Greenson J, Winter J, Smith M, Rogers SJ and Webb SJ (2012) Early behavioral intervention is associated with normalized brain activity in young children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry* **51**(11), 1150–1159. https://doi.org/10.1016/j.jaac.2012.08.018

Dietterich TG (2000) Ensemble methods in machine learning. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science* (Vol. 1857, pp. 1–15). Springer, Berlin: Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, Montiel-Nava C, Patel V, Paula CS, Wang C, Yasamy MT and Fombonne E (2012) Global prevalence of autism and other pervasive developmental disorders. *Autism Research* **5**(3), 160–179. https://doi.org/10.1002/aur.239

Eslami T, Mirjalili V, Fong A, Laird AR and Saeed F (2019) ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Frontiers in Neuroinformatics* **13**, Article 70. https://doi.org/10.3389/fninf.2019.00070

Esteva A, Kuprel B, Novoa R, Ko J, Swetter SM, Blau HM and Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118. https://doi.org/10.1038/nature21056

Farhat T, Akram S, Rashid M, Jaffar A, Bhatti SM and Iqbal MA (2025) A deep learning-based ensemble for autism spectrum disorder diagnosis using facial images. *PLoS One* **20**(4), e0321697. https://doi.org/10.1371/journal.pone.0321697

Hammond P, Forster-Gibson C, Chudley A, Allanson J, Hutton T, Farrell S, McKenzie J, Holden J and Lewis M (2008) Face–brain asymmetry in autism spectrum disorders. *Molecular Psychiatry* **13**, 614–623. https://doi.org/10.1038/mp.2008.18

Hashemi J, Dawson G, Carpenter KLH, Campbell K, Qiu Q, Espinosa S, Marsan S, Baker JP, Egger HL and Sapiro G (2021) Computer vision analysis for quantification of autism risk behaviors. *IEEE Transactions on Affective Compu* **12**(1), 215–226. https://doi.org/10.1109/TAFFC.2018.2868196

Hashemi J, Qiu Q and Sapiro G (2015) Cross-modality pose-invariant facial expression. In 2015 *IEEE International Conference on Image Processing (ICIP)* (pp. 4007-4011). IEEE. Available at https://doi.org/10.1109/ICIP.2015.7351558

He K, Zhang X, Ren S and Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. Available at https://doi.org/10.1109/CVPR.2016.90

Jaiswal M (2020) Interpreting Multimodal Machine Learning Models Trained for Emotion Recognition to Address Robustness and Privacy Concerns, 13716-13717. Available at https://doi.org/10.1609/aaai.v34i10.7130

Khan SA and Adarsh MJ (2025) Deep learning-based system for early screening and assessment of autism spectrum disorder. *International Research Journal of Modernization in Engineering, Technology and Science* **7**(9), 1–6. https://doi.org/10.56726/IRJMETS82502

Kiela D, Bhooshan S, Firooz H, Perez E and Testuggine D (2020) Supervised multimodal bitransformers for classifying images and text. Available at https://doi.org/10.48550/arXiv.1909.02950 (accessed 16 February 2026).

Li Y, Huang W-C and Song P-H (2023) A face image classification method of autistic children based on the two-phase transfer learning. *Frontiers in Psychology* **14**, 1226470. https://doi.org/10.3389/fpsyg.2023.1226470

Litjens G, Kooi T, Ehteshami Bejnordi B, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B and Sánchez CI (2017) A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lord C, Rutter M, DiLavore PC, Risi S, Gotham K and Bishop S (2012) *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. Torrance, CA: Western Psychological Services.

McLaughlin CS, Grosman HE, Guillory SB, Isenstein EL, Wilkinson E, Trelles MDP, Halpern DB, Siper PM, Kolevzon A, Buxbaum JD, Wang AT and Foss-Feig JH (2021) Reduced engagement of visual attention in children with autism spectrum disorder. *Autism* **25**(7), 2064–2073. https://doi.org/10.1177/13623613211010072

Mehrabi N, Morstatter F, Saxena N, Lerman K and Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys* **54**(6), Article 115. https://doi.org/10.1145/3457607

Rane NL, Mallick SK, Kaya O and Rane J (2024) Explainable and trustworthy artificial intelligence, machine learning, and deep learning. In *Applied Machine Learning and Deep Learning: Architectures and Techniques*. Deep Science Publishing, pp. 167–191. https://doi.org/10.70593/978-81-981271-4-3_9

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 618-626). IEEE. Available at https://doi.org/10.1109/ICCV.2017.74 (accessed 22 February 2026).

Sewani H and Kashef R (2020) An autoencoder-based deep learning classifier for efficient diagnosis of autism. *Children* **7**(10), 182. https://doi.org/10.3390/children7100182

Sousa A, Young K, D'Aquin M, Zarrouk M and Holloway J (2023) Introducing CALMED: multimodal annotated dataset for emotion detection in children with autism. Available at https://doi.org/10.48550/arXiv.2307.13706. arXiv preprint arXiv: 2307.13706.

Sun Q, Akman A and Schuller BW (2024) Explainable artificial intelligence for medical applications: A review. Available at https://doi.org/10.48550/arXiv.2412.01829.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698

Tan M and Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)* (pp. 6105-6114). PMLR. Available at https://proceedings.mlr.press/v97/tan19a.html (accessed 8 February 2026).

Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane PA and Wong TY (2019) Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology* **103**(2), 167–175. https://doi.org/10.1136/bjophthalmol-2018-313173

Topol E (2019) *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY, USA: Basic Books.

Wang S, Jiang M, Duchesne XM, Laugeson EA, Kennedy DP, Adolphs R and Zhao Q (2015) Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron* **88**(3), 604–616. https://doi.org/10.1016/j.neuron.2015.09.042

Wei Q, Dong W, Yu D, Wang K, Yang T, Xiao Y, Long D, Xiong H, Chen J, Xu X and Li T (2024) Early identification of autism spectrum disorder based on machine learning with eye-tracking data. *Journal of Affective Disorders* **358**, 326–334. https://doi.org/10.1016/j.jad.2024.04.049

Xia Q, Zheng H, Zou H, Luo D, Tang H, Li L and Jiang B (2025) A comprehensive review of deep learning for medical image segmentation. *Neurocomputing* **613**, 128740. https://doi.org/10.1016/j.neucom.2024.128740

Zwaigenbaum L, Bauman ML, Fein D, Pierce K, Buie T, Davis PA, Newschaffer C, Robins DL, Wetherby AM, Choueiri R, Kasari C, Stone WL, Yirmiya N, Estes A, Hansen RL, McPartland JC, Ratowicz MR, Carter A, Granpeesheh D, Mailloux Z, Smith Roley S and Wagner S (2015) Early screening of autism spectrum disorder: Recommendations for practice and research. *Pediatrics* **136**(Supplement 1), S41–S59. https://doi.org/10.1542/peds.2014-3667D

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 19:36:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1463423626101698