Original Paper

Balancing Human and AI Features for Digital Autism Diagnostics from Home Videos: Human-In-The-Loop Machine Learning Study

Abdur Rasool[1], PhD; Kaitlyn Dunlap[2], BA; Arman Husic[2], BSc; Aaron Kline[2], BS; Parnian Azizian[2], BSc; Dennis P Wall[2*], PhD; Peter Washington[3*], PhD

[1]Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI, United States

[2]Department of Psychiatry and Behavioral Sciences (by courtesy), Stanford University, California, CA, United States

[3]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States

[*]these authors contributed equally

Corresponding Author:
Peter Washington, PhD
Department of Medicine
University of California, San Francisco
500 Parnassus Avenue
San Francisco, CA, 94143
United States
Phone: 1 (415) 353 2067
Email: peter.washington@ucsf.edu

Abstract

Background: Autism spectrum disorder (ASD or autism) diagnosis faces significant delays due to specialist shortages and subjective, clinic-based assessments. Although AI offers a potentially scalable solution via automated video analysis, current state-of-the-art deep learning models often fail to capture subtle, clinically nuanced social behaviors essential for reliable detection.

Objective: This study aimed to develop and evaluate a human-in-the-loop (HITL) framework that combines automated AI feature extraction with crowdsourced human annotation for autism classification on naturalistic home videos. We targeted low-confidence AI facial feature extraction outputs to be replaced by human features. We also sought to quantify the relationship between the confidence threshold used for swapping AI features with human features, the volume of human replacement of AI features, and the predictive diagnostic performance.

Methods: We implemented a HITL pipeline integrating deep learning models for facial emotion (7-class) feature extraction. A confidence threshold applied to the emitted probability of the AI feature extractor model was used to identify low-confidence facial features or those that the AI model was unable to make a prediction for (henceforth referred to as "unresolved"), which were then replaced by human "feature extraction" (ie, annotation) conducted by rigorously vetted crowd workers. We evaluated this approach on 100 high-quality videos (50 autism and 50 neurotypical) from the GuessWhat? digital therapeutic app. Performance was assessed across decision thresholds (0.1-0.9 in increments of 0.1) using sensitivity, specificity, F_1-score, and area under the receiver operating characteristic curve (AUROC), with 5-fold cross-validation.

Results: Across decision thresholds, performance improved overall but was nonmonotonic at intermediate thresholds. At a threshold of 0.9, 59.7% (597/1000) of frames were replaced by human annotations, and mean sensitivity increased from 0.51 (SD 0.24) at a threshold of 0.1 to 0.66 (SD 0.10) at a threshold of 0.9, an absolute increase of approximately 15 percentage points. The HITL-enhanced multimodal model achieved an aggregate out-of-fold AUROC of 0.76, compared with 0.54 for the AI-only model.

Conclusions: In this proof-of-concept study, replacing AI features with human features tended to be associated with higher observed autism classification performance in this particular dataset. We targeted low-confidence and unresolved frames for human annotation. We note that these findings are exploratory and thus do not isolate the independent causal effect of human replacement or establish clinical diagnostic validity.

(JMIR AI 2026;5:e84699) doi: 10.2196/84699

XSL•FO
RenderX

KEYWORDS

autism spectrum disorder; human-in-the-loop AI; artificial intelligence; digital phenotyping; video analysis

Introduction

Autism spectrum disorder (ASD; also known as autism) is a neurodevelopmental condition characterized by social interaction challenges, communication difficulties, and restricted and repetitive behaviors, typically identifiable in early childhood [1,2]. Many children do not receive a formal diagnosis until the age of 4 years or later due to limited specialist availability, long wait times, and reliance on subjective clinical observations and caregiver reports [3]. In the United States, autism prevalence has risen from 1 in 150 children in 2000 to 1 in 36 by 2020, and yet more than 83% of counties lack sufficient diagnostic services [3-5]. These delays hinder early intervention, which improves long-term developmental outcomes and quality of life for individuals with autism [2,6-9].

With increasing autism prevalence and diagnostic resource shortages, scalable and accessible diagnostic tools are urgently needed [1]. Digital phenotyping, which uses mobile health technologies and wearables to quantify human behavior, enables real-world behavioral data collection and AI-driven analysis [2,10,11]. Mobile apps allow structured video capture of children engaging in interactive tasks while AI models analyze behavioral patterns associated with autism [10,12]. A notable example is a Google Glass–based system that provides real-time feedback on emotional expressions, aiding both data collection and social engagement. These digital tools demonstrate how scalable alternatives to in-clinic evaluations can support early diagnosis and treatment [2,7].

AI-based digital phenotyping approaches can and will benefit from a sharper focus on improving their ability to capture nuanced social behaviors critical for autism diagnosis [2,13-15]. Key autism-related indicators, such as fleeting eye contact or microexpressions, are frequently misclassified by AI [2,16,17]. Subtle social cues vary across children and contexts, making them difficult to classify. AI models are also prone to overfitting, particularly when trained on limited datasets, leading to poor generalization to new cases.

Human-in-the-loop (HITL) approaches address these challenges by integrating human expertise into AI pipelines [18,19]. Human annotators can more reliably label complex behaviors that AI alone may misinterpret [20]. Prior studies [21,22] show that trained nonexpert crowd workers can reliably annotate sociobehavioral features relevant to autism classification. Human annotators often detect subtle cues that AI models overlook. Hybrid approaches combining AI with human annotations have the potential to enhance classification reliability by reducing noise, improving feature granularity, and increasing model interpretability, particularly in diverse, real-world environments [1,20,23-25].

We introduce a HITL AI framework to overcome the critical limitation of automation failures when extracting nuanced social cues for autism diagnosis. We used deep learning models for automated feature extraction via a facial emotion feature extraction (FEFE) model using a pretrained convolutional network to classify 7 emotions and an eye gaze (EG) model to quantify eye contact. We systematically identified low-confidence or unclassifiable (usually due to the face not being detectable, henceforth referred to as “unresolved”) FEFE outputs for replacement with human annotations provided by a curated human workforce identified through crowdsourcing campaigns. We evaluated the framework on a curated dataset of 100 high-quality home videos from the GuessWhat? digital therapeutic app, and we compared the performance of autism classification when using (1) purely automated feature extraction and (2) features extracted by a curated pool of nonexpert humans.

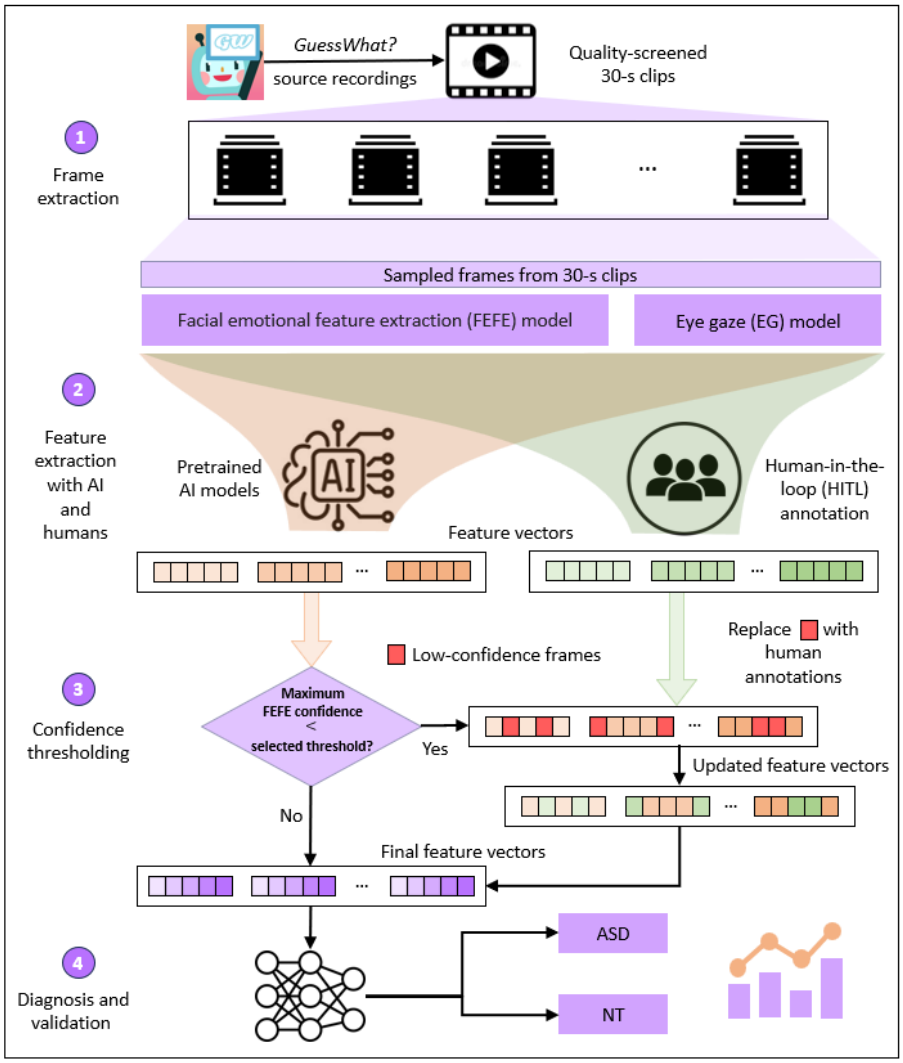
Methods

Overview

We designed a framework (Figure 1) that analyzes home videos using AI models (FEFE and EG) to extract facial emotion and eye-gaze features. We implemented a HITL pipeline integrating deep learning models for FEFE across 7 classes. A confidence threshold applied to the emitted probability of the AI feature extractor model was used to identify low-confidence facial features or those for which the AI model was unable to make a prediction (henceforth referred to as “unresolved”), which were then replaced by human “feature extraction” (ie, annotation) conducted by rigorously vetted crowd workers. We then quantified the relationship between the facial confidence threshold and classification performance. Details are provided in the next subsections.

XSL•FO
RenderX

Figure 1. Framework of an AI-enhanced, human-in-the-loop methodology for autism classification. AI models initially extract facial emotion and gaze features from video frames. When the confidence score is below a prespecified decision threshold, the AI-based features are replaced by human "feature extraction" (ie, annotation). A neural network then classifies autism (autism spectrum disorder) or neurotypical using a combination of AI-extracted and human-annotated features. ASD: autism spectrum disorder; FEFE: facial emotion feature extraction; HITL: human-in-the-loop: NT: neurotypical.



GuessWhat? Data Acquisition

We used data [10,22,26] collected from a cross-platform mobile game called GuessWhat? This game is designed to encourage prosocial play and interaction between children and their parents in a home environment. The GuessWhat? game included 2 primary activities: charades and quizzes. In charades, children acted out emotions, characters, and objects, while in quizzes, they answered questions about colors, shapes, and word spellings. During each session, parents acted as mediators and recorded their child's gameplay by placing the smartphone on their forehead, ensuring that the front-facing camera remained oriented toward the child (Figure S1 in Multimedia Appendix 1). This setup provided a consistent frame with the child's face visible, maintaining a distance of approximately 60 to 180 cm from the camera. At the end of each 90-second session, parents were given the option to review and share the video with the research team at Stanford University [10,13,27].

The collected dataset included videos from children diagnosed with autism and neurotypical children. These videos were sourced remotely, ensuring a diverse dataset that covers various demographic backgrounds, as provided by caregivers during the registration process. Diagnosis was established through caregiver report and subsequent consensus confirmation by 2 independent clinical reviewers who were not involved in the development of the AI models or the HITL framework. These reviewers assessed the provided video data independently of the study's analytical pipeline against standard developmental benchmarks, though formal Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) or Autism Diagnostic Interview–Revised (ADI-R) assessments [28] were not conducted specifically for this video subset. Videos were screened to exclude any with poor lighting, corrupted files, or irrelevant content to ensure high-quality, analyzable footage and maintain input consistency for AI model processing.

To streamline data processing, each video was uploaded directly to an Amazon S3 bucket through an automated process within the app. The standardized format of each gameplay session, where the device was held in landscape mode and the child remained visible in the frame, ensured consistency across the dataset. Any videos not meeting these standards were excluded from the study.

XSL•FO
RenderX

By using a smartphone, this approach lowers barriers to health care, making it feasible for a larger population of children to receive care outside of clinical settings. This method is also time- and cost-efficient compared with in-lab assessments, and recording in a home environment promotes more naturalistic behavior in children, potentially yielding data that more accurately represent real-world interactions.

Parents or guardians reviewed each recording and provided consent for research use; the analysis subset was selected from recordings for which consent also permitted public sharing. From the selected recordings, deidentified uncropped still frames were placed in a private repository and made available only to individual authorized Microworkers participants through task-specific secure URLs. Workers viewed isolated frames only and had no access to source videos, participant identities, diagnostic labels, or broader video context.

The initial dataset contained a total of 1093 raw video recordings (Table 1), comprising 985 videos of children with autism and 108 videos of neurotypical children. This dataset was divided into autism and neurotypical categories, with ages ranging from 3 to 12 years. The demographic breakdown, as shown in Table 1, includes gender distribution and age groups, highlighting a comprehensive spread across age and gender for both autism and neurotypical groups.

Table 1. Demographic and structural overview of datasets, showing the initial source dataset (n=1093 videos) and the manually filtered high-quality analysis dataset of unique children (n=100; 50 autism, 50 neurotypical) used for final analysis.

Datasets and characteristics	Autism	Neurotypical	Total
Initial source dataset			
Number of videos	985	108	1093
Size (GB)	33.11	4.85	37.96
Number of boys	879	69	948
Number of girls	106	39	145
Age range (years)	3-12	3-12	3-12
Age group distribution (years), n			
3-5	565	30	595
6-10	263	48	311
≥11	157	30	187
Filtered high-quality analysis dataset			
Unique individuals and videos, n	50	50	100
Number of boys	25	26	51
Number of girls	25	24	49
Age range (years)	3-12	3-12	3-12
Age group distribution (years), n			
3-5	17	14	31
6-10	15	27	42
≥11	18	9	27

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the Stanford University Institutional Review Board (protocol 39562; approved March 21, 2023). The study team confirmed that Protocol 39562 and the associated parental consent process covered the restricted third-party crowdsourcing workflow used here. Parents or guardians provided informed consent for research use of the recordings, and the analyzed subset was selected from recordings for which consent also permitted public sharing. However, the source videos were not publicly released or provided to crowd workers. For crowdsourcing, individual deidentified, uncropped still frames sampled from the analyzed 30-second clips were placed in a private repository and made available only to authorized Microworkers participants through task-specific secure URLs. Workers had no access to source videos, participant identities, diagnostic labels, or a downloadable public image repository.

Manual Filtration of High-Quality Videos

To create a dataset suitable for evaluation, we manually screened the recordings for analyzable facial visibility and framing. For each retained recording, one contiguous 30-second segment was selected in which the target child's face was sufficiently large, clear, and in focus, and no other face was the primary visible subject. Segments with persistent blur, severe occlusion, poor lighting, or an unsuitable dominant subject were excluded. For children with multiple recordings, only recordings meeting these criteria were included. Ten evenly spaced still frames were subsequently sampled from each selected segment for feature extraction. The curated dataset comprised 50 videos from unique children with autism and 50 videos from neurotypical children.

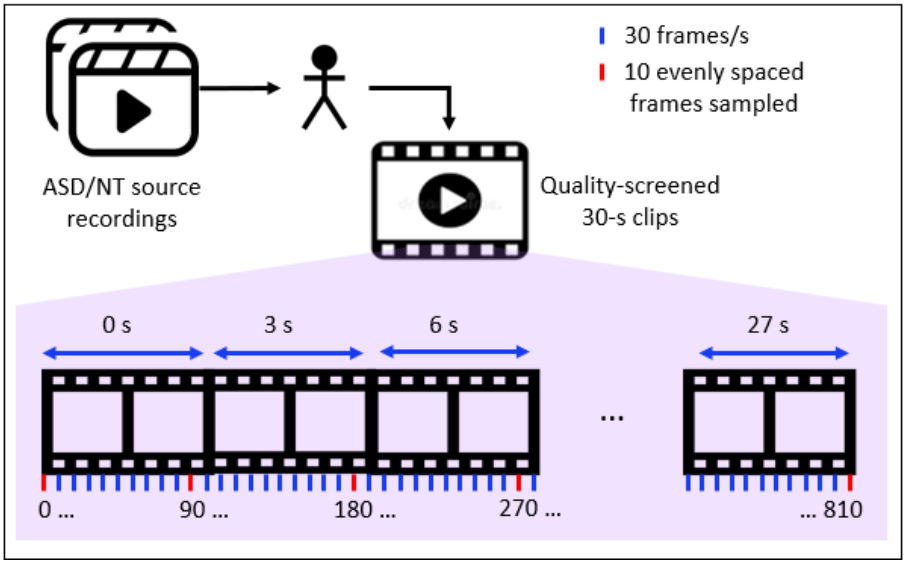
XSL•FO
RenderX

The autism subset included 25 boys and 25 girls aged 3 to 12 years, while the neurotypical subset included 26 boys and 24 girls with a similar age distribution. Further demographic details are summarized in Table 1.

Data Preprocessing and Frame Extraction

Following the manual quality screening procedure described above, one contiguous 30-second segment was selected from each eligible recording. The selected segment prioritized clear, sufficiently large, and in-focus visibility of the target child's face, without another face serving as the primary subject. Ten evenly spaced still frames were then sampled from each selected segment for FEFE and EG feature extraction. First, the total number of frames F_{total}in the video was calculated as the product of the video duration T (in seconds) and its frame rate fps. To illustrate (Figure 2), consider a 30-second video clip recorded at 30 frames per second, which would yield a total frame count as follows:

Figure 2. Schematic of the systematic frame sampling procedure for feature extraction from home videos in the GuessWhat? autism study, illustrating the calculation of step value for selecting evenly spaced frames across a video's duration. ASD: autism spectrum disorder; NT: neurotypical.



$$F_{total} = T \times fps = 30\ sec \times 30\ fps = 900 \text{ frames} \quad (1)$$

To extract a representative sample of N frames (eg, 10 frames) from this video, a step value S was computed by dividing the total frames by the number of frames to be extracted:

$$Step\ Value = \left\lfloor \frac{Total\ number\ of\ frames}{Number\ of\ frames\ to\ extract} \right\rfloor \quad (2)$$

$$S = \left\lfloor \frac{F_{total}}{N} \right\rfloor = \left\lfloor \frac{900}{10} \right\rfloor = 90 \quad (3)$$

This calculated step value S ensured frames were evenly spaced, with every 90th frame selected for extraction in this example. Starting at frame 0, the first frame was extracted; then, proceeding at intervals of 90 frames, subsequent frames were extracted at positions $F_i = i \times S$, where i=0,1,2,…9. Consequently, the extracted frame positions were 0, 90, 180, ..., 810, ensuring that frames were distributed evenly across the video. At 30 frames per second, these corresponded approximately to 0, 3, 6, 9, 12, 15, 18, 21, 24, and 27 seconds. While the specific step value differed for videos with different frame rates, the calculation method remained consistent. This systematic approach guaranteed a balanced and representative sample for facial expression tasks and minimized potential bias from frame selection, regardless of the source video's frame rate.

Sampled still frames were passed directly to the modality-specific feature extractors. The FEFE extractor passed each frame to the fer Python package, which returned scores for angry, disgust, fear, happy, sad, surprise, and neutral when a face or emotion result was available; when no result was returned, the implementation stored zeros for all 7 facial features. The EG extractor processed the same still frames using MediaPipe Face Mesh. The resulting frame-level features were standardized within each cross-validation training fold before classifier training.

FEFE Model

Our facial expression analysis involved the classification of 7 emotions (angry, disgust, fear, happy, sad, surprise, and neutral) using a pretrained facial expression recognition (FER) model implemented via the *fer* Python package [29]. This model was originally trained on the FER-2013 dataset containing 35,887 grayscale images [30]. We used this basic emotion model as an initial, interpretable proxy for social engagement and affective expression, which are core domains impaired in autism. Our choice of the 7 emotion categories was grounded in the Ekman foundational theoretical framework of universal basic emotions, providing a clinically intuitive and widely recognized starting point for automated screening [31]. While action unit (AU) analysis provides more granular facial movement data, emotion classifications offer a higher level, clinically intuitive framework for initial automated screening [32]. We acknowledge that applying this model to pediatric faces without fine-tuning represents a limitation, potentially reducing accuracy for child-specific expressions [33].

The preprocessing pipeline included automatic padding, face region normalization, and coordinate offsetting for optimal face framing. The emotion classifier generated probability

XSL•FO
RenderX

distributions across the 7 emotion categories through local model inference, with each frame processed sequentially and results stored for further analysis. Conceptually, the pipeline supports an active learning framework where challenging samples identified through low confidence could be used to iteratively refine the model via human annotations. In this study, the FEFE model was used as a static feature extractor without retraining.

Processed features from each modality were compiled into a unified, column-wise matrix for each video, creating a comprehensive dataset for robust multiclass classification. This approach ensured high-quality feature extraction while maintaining the flexibility to integrate human annotations when model confidence was low.

EG Model

As a secondary exploratory analysis, we used MediaPipe Face Mesh to derive a binary gaze feature from eye-corner geometry. For each detected face, the implementation used 4 eye-corner landmarks (33, 133, 362, and 263) to calculate left and right eye-vector angles with atan2. The absolute value of the mean of the 2 eye-vector angles was compared with a fixed threshold of 8 degrees. No eye aspect ratio, blink detection, or eye symmetry calculation was used. The stored EG feature was coded as eye contact=1 and no eye contact=0. For frames with detected facial landmarks, the implementation assigned 1 when the average absolute eye-vector angle G was below 8 degrees and assigned 0 when G≥8°. When Face Mesh did not detect facial landmarks, the implementation returned NaN, representing a missing numeric value rather than either gaze class. The full gaze-angle calculation and gaze-humanization procedure are provided in section S1.1 in Multimedia Appendix 1.

Human Annotation Process Using Crowdsourcing

To obtain high-quality human annotations, we designed and conducted "human selection" campaigns on Microworkers.com, ensuring a systematic approach to select the best-performing workers and generate reliable data annotations.

Initially, we developed a strategy for selecting top-performing workers. We extracted 100 frames for emotion and gaze features from 10 high-quality GuessWhat? videos. For the pilot screening campaigns, we targeted approximately 100 workers per modality from the Microworkers international zone who met the specified platform performance rating criterion. Three top-performing workers were subsequently selected for the production annotation task. Crowdsourced annotators were specifically tasked with labeling discrete behavioral features, namely the presence of 1 of 7 emotion states and 1 of 2 EG classes, based on provided video frames. For the emotion annotation task, annotators were also provided with explicit criteria for the "no or partial face" (NPF) class: a frame should be labeled NPF if (1) the face was fully occluded, (2) less than 50% of the facial region was visible, or (3) the face was severely out-of-focus or poorly illuminated such that key landmarks could not be discerned. The gaze task used only binary classification (eye contact or no eye contact) without an NPF class. These selected samples covered diverse feature sets, enabling the evaluation of worker performance across varying feature vectors. Simultaneously, our research team collaboratively annotated these frames, with final labels verified by a senior domain expert, to establish a robust ground-truth benchmark. The selected frames were stored in a private repository and made available only to authorized Microworkers participants through task-specific secure URLs included in the task templates. These were uncropped, deidentified still frames sampled from the analyzed clips, not face-cropped images and not video clips. Each worker viewed individual frames only and did not receive access to the source video or its broader temporal context.

To further protect privacy, access to the frames was restricted to the task interface via secure URLs, with no downloadable repository available. All crowd workers were bound by the Microworkers.com platform's terms of service, which prohibit data misuse. Importantly, annotators had no access to participant identities, diagnostic labels, or full video context, limiting their task to labeling discrete behavioral features from isolated frames. Separate templates were created for each modality to simplify tasks for crowd workers. Workers were provided with adequate time and reasonable compensation to complete the tasks.

Once the crowdsourced annotations were received, we compared them with the ground truth data to calculate each worker's average performance. The performance evaluation of crowd workers, including accuracy and task completion time, is summarized in Figure S2 in Multimedia Appendix 1, demonstrating the selection of top-performing workers based on their reliability and efficiency. For the production task, the 100-video dataset contained 1000 sampled frames per modality. Each sampled frame had up to 3 annotations from the selected workers. The humanization pipeline derived a valid plurality label across the retained human-labeled entries from the available annotations. For facial features, a valid human plurality label changed the stored AI feature when the maximum AI probability was below the selected facial confidence threshold and the human label differed from the AI winning class. For gaze features, a valid human plurality label changed the stored binary AI value only when it differed from the AI value. Detailed step-by-step examples of the confidence-based replacement process, including visual illustrations of feature vector transformations across thresholds, are provided in section S1.2 and Table S1 in Multimedia Appendix 1. This process ensured that the annotations used in our study were accurate and reliable, ultimately enhancing the quality of the data for our research objectives.

Confidence-Based HITL Annotation Replacement Process

For FEFE, operational HITL decisions were based on maximum confidence, defined as the highest predicted probability among the 7 FEFE emotion classes for a given frame. When the stored FEFE probabilities were all zero, no AI winning class existed, and the row was treated as an unresolved AI output. For a nonzero row whose maximum probability was greater than or equal to the selected threshold, the AI winning class and its maximum probability were retained, while nonwinning class values were set to zero. For a nonzero row below the threshold, the corresponding human facial annotations were retrieved and normalized before voting. If the valid human plurality label differed from the AI winning class, the row was replaced by a

XSL•FO
RenderX

7-class 1-hot human vector; if the labels agreed, the existing AI row was retained. For an all-zero row, any valid human plurality label replaced the unresolved row with a 7-class 1-hot vector; if no valid human label was available, the all-zero row was retained.

For the AI-only and HITL FEFE inputs, NPF was used as an annotation category and in descriptive class-mass analysis, but was not included as an eighth model-input feature. These FEFE inputs therefore contained 7 emotion features, and the corresponding multimodal inputs contained 8 columns: 7 FEFE features plus 1 binary EG feature. The separate pure-human comparator retained NPF as an eighth facial feature; consequently, its facial input was 8-dimensional, and its combined facial-plus-gaze input was 9-dimensional.

Gaze humanization was performed separately from the facial confidence-threshold procedure. The EG feature was coded as eye contact=1 and no eye contact=0 and did not contain an NPF category. For each frame, valid human gaze labels were collected, and the majority label was selected; in a tie, no eye contact was selected. When the valid human plurality label differed from the stored AI binary value, the value was replaced with 1 for eye contact or 0 for no eye contact. When the labels agreed, or when no valid human gaze annotation was available, the stored AI value was retained.

The FEFE human-intervention ratio (R_{HITL} (τ)) was calculated as:

$$R_{\mathrm{HITL}}(\tau) = \frac{N_{\mathrm{zero,human}} + N_{\mathrm{nonzero,diff}}(\tau)}{N_{\mathrm{frames}}} \times 100\% \qquad (4)$$

Where $N_{\mathrm{zero,human}}$ is the number of all-zero FEFE rows that received a valid human plurality label, and $N_{\mathrm{nonzero,diff}}$ (τ) is the number of nonzero AI rows below threshold τ whose valid human plurality label differed from the AI winning class. N_{frames} is the total number of analyzed frames.

The reported FEFE humanization ratio counted valid humanization of either unresolved all-zero FEFE rows or nonzero low-confidence rows with a differing human label. It excluded rows with no valid human label, NPF responses, same-label human annotations, and retained AI rows. It did not count every numeric row change, because high-confidence AI rows were also sparsified by retaining only the winning class. The EG replacement ratio was calculated as the percentage of binary AI values changed after valid binary gaze voting with the prespecified tie rule. The complete FEFE and EG workflows are provided in section S1.2 and Table S1 in Multimedia Appendix 1.

Feature Aggregation, Model Training, and Evaluation

The final classification of each video (and thus each participant) was performed by aggregating frame-level features across the entire video. Following the systematic sampling procedure detailed in the Data Preprocessing and Frame Extraction section, we extracted n=10 evenly spaced frames from each 30-second video. For each frame, we obtained a feature vector: for FEFE, this was a 7-dimensional probability vector, with valid low-confidence human replacements encoded as 7-dimensional 1-hot vectors; for EG, this was a 1-dimensional binary feature coded as 1 for eye contact and 0 for no eye contact. The sequence was processed by a 2-layer bidirectional long short-term memory (LSTM) with 64 hidden units per direction and dropout of 0.3. The implementation used the output at the final sequence position as the input to a fully connected sigmoid classifier. At this position, the forward-direction component reflects processing through the sequence, whereas the backward-direction component is the backward activation aligned with the final input position and is not, by itself, a conventional full-sequence reverse summary. The resulting output was used to generate 1 video-level autism-versus-neurotypical prediction. All reported performance metrics (accuracy, F_1-score, etc) were calculated based on these video-level predictions.

For model evaluation, we used 5-fold cross-validation over the full 100-video analysis set. The 5 folds were created using shuffled k-fold cross-validation with a fixed random seed of 42. In each fold, 80 videos were used for model training, and the remaining 20 videos were held out for evaluation. Each video corresponded to 1 participant, and each participant contributed 1 held-out prediction across the complete cross-validation procedure. Feature standardization was fitted using the training data within each fold and then applied to the corresponding held-out data. A newly initialized bidirectional LSTM was trained separately in each fold for 50 epochs using Adam optimization with a learning rate of 0.001, binary cross-entropy loss, and a batch size of 16. Sigmoid outputs were converted to binary video-level predictions using a cutoff of 0.5; this classification cutoff was distinct from the facial HITL confidence thresholds of 0.1-0.9. Fold-level performance was summarized as the mean and SD of the 5 held-out fold metrics; thus, the mean (SD) values reported in subsections of Results and Figures S3, S4, and S6 in Multimedia Appendix 1 represent between-fold variability, not within-fold bootstrap variability.

For aggregate out-of-fold evaluation, the held-out predictions from all 5 folds were concatenated so that the final evaluation set contained 1 prediction for each of the 100 videos. Aggregate performance metrics, the CIs, and aggregate confusion matrices shown in the last subsections of Results were calculated from this pooled out-of-fold set, not from a separate isolated 20-video test set. Wilson score 95% CIs were calculated for accuracy, sensitivity, and specificity. Percentile bootstrap 95% CIs, based on 2000 resamples with a fixed random seed of 42, were calculated for F_1-score and area under the receiver operating characteristic curve (AUROC).

Results

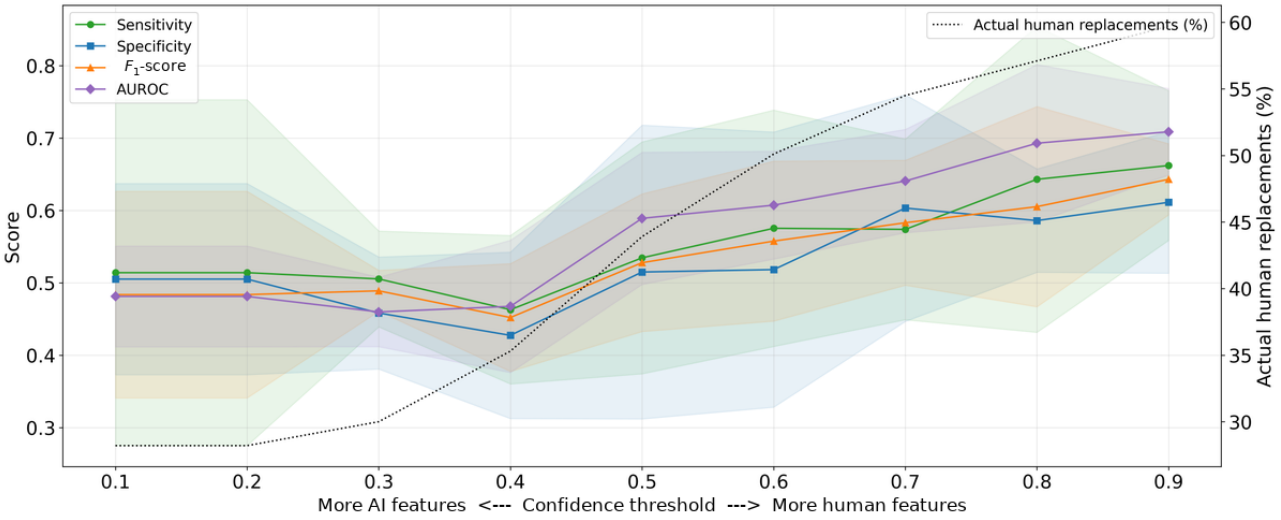
Threshold-Dependent Performance of HITL Framework

Overall, performance was higher at the upper end of the facial confidence-threshold sweep, although the relationship was nonmonotonic and did not improve over the AI-only baseline at every threshold. We evaluated unimodal autism-classification models using both FEFE and EG features. Because the FEFE configuration showed the stronger observed performance, the facial confidence-threshold sweep (Figure 3 and Figures S3 and

XSL•FO
RenderX

S4 in Multimedia Appendix 1) was our primary focus. We analyzed the performance of the FEFE model (Figure 3) across confidence thresholds (threshold 0.1-0.9) to evaluate the impact of progressively integrating HITL annotations. The valid FEFE humanization ratio, defined as the proportion of frames receiving a valid human replacement under the rule above, was 28.2% at thresholds of 0.1 and 0.2 and increased to 59.7% at a threshold of 0.9 (Table S2 in Multimedia Appendix 1).

Figure 3. Facial emotion feature extraction performance across facial human-in-the-loop confidence thresholds (threshold 0.1-0.9). Lines show 5-fold mean sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve; shaded bands represent (1 SD) across folds. The dotted line shows the proportion of frames receiving a valid human facial emotion feature extraction replacement. Performance improved overall from a threshold of 0.1 to a threshold of 0.9, although intermediate values were nonmonotonic. AUROC: area under the receiver operating characteristic curve.



The mean (SD) F_1-score across 5 cross-validation folds increased from 0.484 (0.143) at a threshold of 0.1 to 0.643 (0.049) at a threshold of 0.9. Mean (SD) sensitivity increased from 0.514 (0.239) to 0.662 (0.103), mean specificity from 0.505 (0.132) to 0.611 (0.098), and mean AUROC from 0.481 (0.070) to 0.709 (0.060). F_1-score and AUROC were highest at a threshold of 0.9 among the evaluated thresholds. The temporary decrease in F_1-score at a threshold of 0.4 (mean 0.452, SD 0.074) shows that increasing the confidence threshold does not guarantee an improvement at each step, though we believe this is simply an artifact of our dataset.

Precision-recall curves for the FEFE model are provided in Figure S5 in Multimedia Appendix 1. Out-of-fold average precision increased overall from 0.57 at a threshold of 0.1 to 0.72 at a threshold of 0.9, with the highest value observed at a threshold of 0.8. Because a threshold of 0.9 was selected after inspecting the threshold-sweep results obtained from the same 5-fold cross-validation data, this choice should be regarded as exploratory rather than as an independently validated operating point. The reported performance may therefore be optimistically biased because no nested cross-validation or independent tuning set was used.

Analysis of Crowdsourced Annotations and Intervention Rates

Figure 4A summarizes class-level feature distributions across the 1000 sampled facial frames. The AI bars represent summed positive FEFE probability mass, whereas the human bars represent weighted human-label mass, including the NPF category. These are feature-mass summaries rather than equivalent frame counts. AI probability mass was highest for happy (229.9) and neutral (167.4). Weighted human-label mass was highest for happy (317.4), neutral (247.9), and NPF (178.6), indicating a broader and more distributed human-label profile.

XSL•FO
RenderX

Figure 4. Class-level AI and weighted human feature distributions and threshold-dependent human-label replacement. Values are sums of fractions over 1000 frames. Each frame adds 1 to the human bars. Each frame adds 1 to the AI bars if a face was detected, and 0 if no face was detected. (A) AI bars show summed positive facial emotion feature extraction probability mass; human bars show weighted annotation mass, including the no or partial face category. (B) Blue bars show frequencies of positive AI emotion-class scores, including the <0.1 bin; red values show the percentage of frames receiving a valid human facial emotion feature extraction replacement at each confidence threshold.

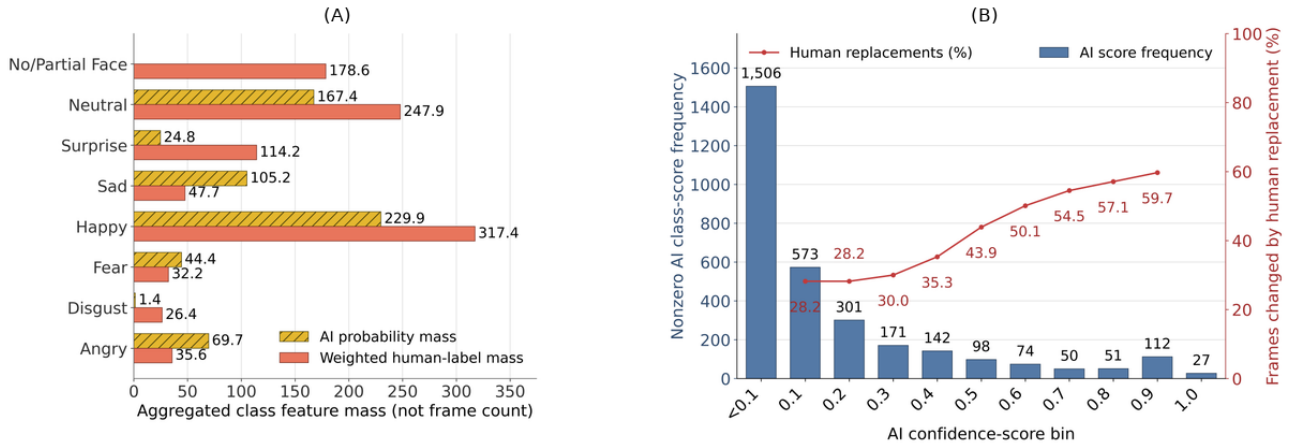


Figure 4B presents 2 separate quantities. The blue bars show the frequency of positive AI emotion-class scores in confidence bins; the <0.1 bin includes positive scores below 0.1. Because 1 frame may contribute several positive class scores, these frequencies are not frame counts and sum to 3105 rather than 1000. The red line shows the percentage of frames receiving a valid human FEFE replacement under each threshold, including valid human labels assigned to unresolved all-zero AI rows. The actual replacement percentages increased from 28.2% at a threshold of 0.1 (and threshold of 0.2) to 59.7% at a threshold of 0.9. At a threshold of 0.9, 597 of 1000 frames received a valid human replacement under this definition. The remaining frames did not receive a different human label.

For the separate EG workflow, gaze humanization changed 483 of the 1000 stored binary gaze values (48.3%): 206 frames changed from no eye contact to eye contact, and 277 changed from eye contact to no eye contact. The AI-only gaze features contained 758 eye contact values and 242 no eye contact values; after gaze humanization, these counts were 687 and 313, respectively.

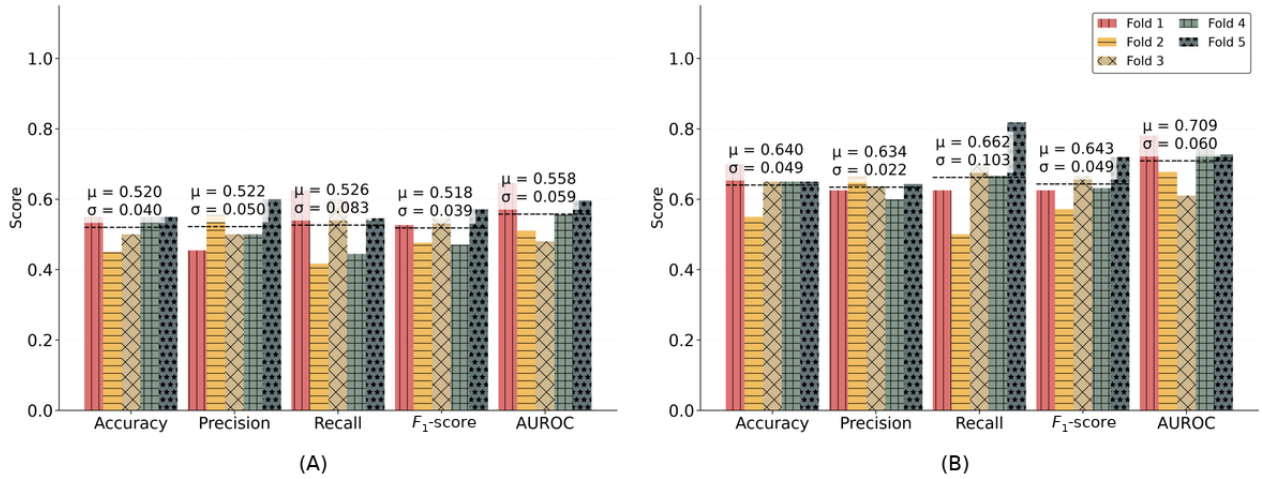
Pure human annotations for FEFE alone (Figure S6A in Multimedia Appendix 1) and combined FEFE with EG (Figure S6B in Multimedia Appendix 1) achieved F_1-scores of 0.593 (0.150) and 0.672 (0.078), respectively, with the combined annotations showing a 48.3% reduction in fold-level F_1-score variability compared with pure FEFE, calculated from the unrounded fold-level SDs.

Performance of Unimodal Feature Extraction

The comparison between the AI-only and HITL FEFE models showed higher mean performance for the HITL configuration across all 5 reported metrics. Figure 5 presents the HITL FEFE model with a facial confidence threshold of 0.9. The mean (SD) F_1-score increased from 0.518 (0.039) to 0.643 (0.049), corresponding to a 24.1% relative gain; the SD increased by 25.6%. Mean (SD) recall (sensitivity) increased from 0.526 (0.083) to 0.662 (0.103), a 25.8% gain, with a 24.1% increase in SD. Mean accuracy (SD) increased from 0.520 (0.040) to 0.640 (0.049), a 23.1% gain, with a 22.5% increase in SD. Mean (SD) precision increased from 0.522 (0.050) to 0.634 (0.022), a 21.5% gain, while its SD decreased by 56%. Mean (SD) AUROC increased from 0.558 (0.059) to 0.709 (0.060), a 27.1% gain, with a small 1.7% increase in SD. These results show higher mean performance for the HITL FEFE configuration across all 5 reported metrics. However, the SD did not decrease uniformly: precision variability decreased, whereas accuracy, recall, and F_1-score variability increased slightly and AUROC variability remained nearly unchanged.

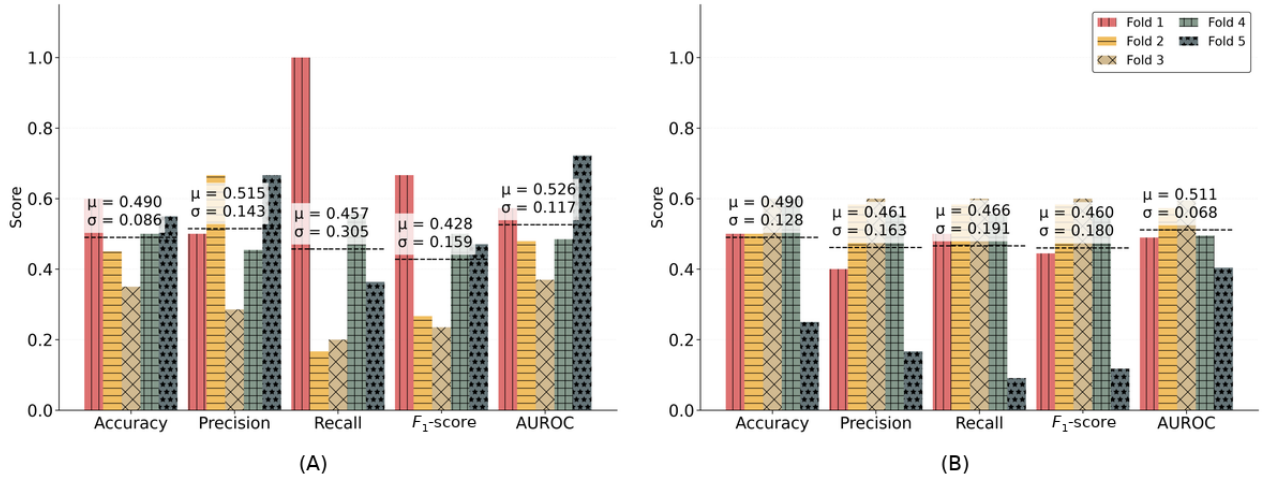
XSL•FO
RenderX

Figure 5. Fold-level mean (SD) performance of facial emotion feature extraction models: (A) AI-only and (B) human-in-the-loop at facial confidence threshold 0.9. The human-in-the-loop configuration showed higher observed mean performance in this exploratory analysis. AUROC: area under the receiver operating characteristic curve.



For the EG model, HITL yielded mixed outcomes with distinct trade-offs (Figure 6). The mean (SD) F_1-score increased slightly from 0.428 (0.159) to 0.460 (0.180), an approximate 7.5% gain; however, its SD increased by approximately 13.2%. Mean (SD) recall increased slightly from 0.457 (0.305) to 0.466 (0.191), an approximate 2% gain, while its SD decreased by approximately 37.4%. Accuracy remained unchanged at 0.490, increasing in variability from 0.086 to 0.128 (approximately 48.8%). Mean (SD) precision decreased from 0.515 (0.143) to 0.461 (0.163), an approximate 10.5% reduction, while its SD increased by approximately 14.0%. Mean (SD) AUROC decreased from 0.526 (0.117) to 0.511 (0.068), an approximate 2.9% reduction, while its SD decreased by approximately 41.9%. These results indicate mixed effects of gaze humanization: small gains in mean F_1-score and recall, unchanged accuracy, and lower precision and AUROC.

Figure 6. Performance comparison of eye gaze models (A) without and (B) with the human-in-the-loop approach. Gaze humanization produced mixed results: mean F1-score and recall increased slightly, accuracy was unchanged, and mean precision and area under the receiver operating characteristic curve decreased. AUROC: area under the receiver operating characteristic curve.



Multimodal Feature Fusion Analysis

For the AI-only and HITL multimodal configurations, the model concatenated the 7 FEFE emotion features with 1 binary EG feature, coded as eye contact=1 and no eye contact=0, giving an 8-column frame-level input. The resulting sequence was then processed with a bidirectional LSTM for autism-versus-neurotypical classification. We evaluated matched AI-only and HITL-enhanced multimodal configurations using the same 5-fold cross-validation splits. Figure 7 reports fold-level means and SDs. The HITL-enhanced configuration at a threshold of 0.9 showed higher mean accuracy, F_1-score, sensitivity, precision, and AUROC than the AI-only configuration in the fold-level summary, although the magnitude of improvement varied across metrics and folds. These results indicate an overall fold-level improvement rather than uniformly superior performance in every fold. Aggregate out-of-fold metrics, calculated by concatenating 1 held-out prediction per video across all 5 folds, are reported separately in Figure 8.

XSL•FO
RenderX

Figure 7. Fold-level mean (SD) performance across the same 5 cross-validation folds for multimodal facial emotion feature extraction+eye gaze classification: (A) AI-only and (B) human-in-the-loop–enhanced at a facial confidence threshold of 0.9. Aggregate out-of-fold metrics from the same 5-fold evaluation are reported separately in Figure 8. AUROC: area under the receiver operating characteristic curve.

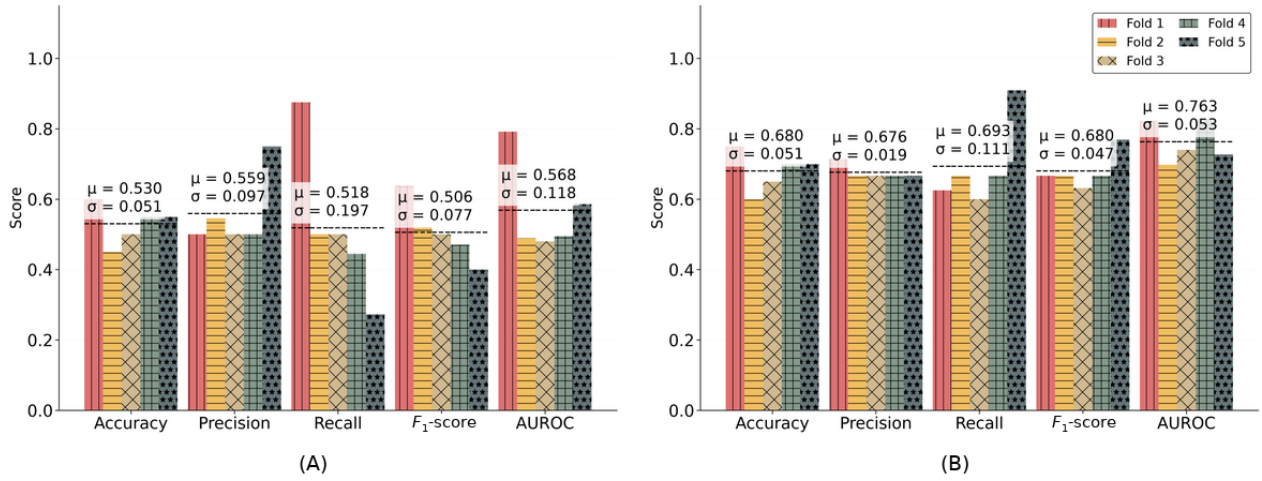
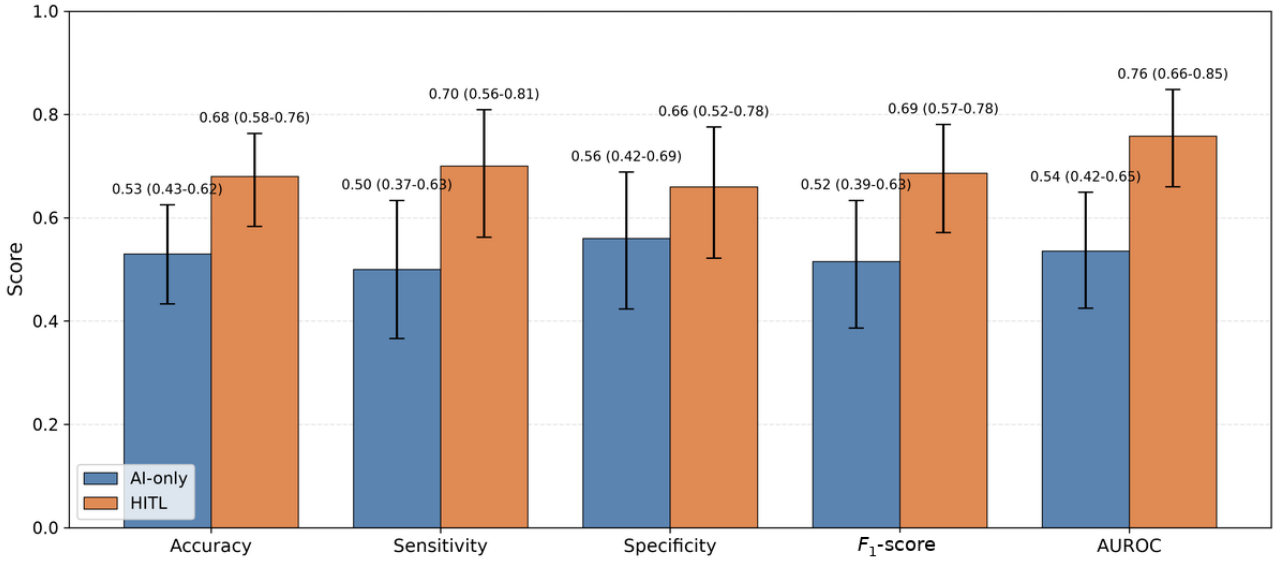


Figure 8. Aggregate out-of-fold performance and 95% CIs for the AI-only and human-in-the-loop–enhanced multimodal models across 100 videos. Human-in-the-loop point estimates were higher for all metrics, with nonoverlapping displayed area under the receiver operating characteristic curve intervals: AI-only, 0.54 (0.42-0.65); human-in-the-loop, 0.76 (0.66-0.85). AUROC: area under the receiver operating characteristic curve; HILT: human-in-the-loop.



The fold-level means also favored HITL for accuracy, F_1-score, and AUROC in every fold. Precision was higher in 4 of 5 folds and recall was higher in 4 of 5 folds. Thus, the results support an overall multimodal HITL improvement, but they should not be described as uniformly better for every metric in every fold.

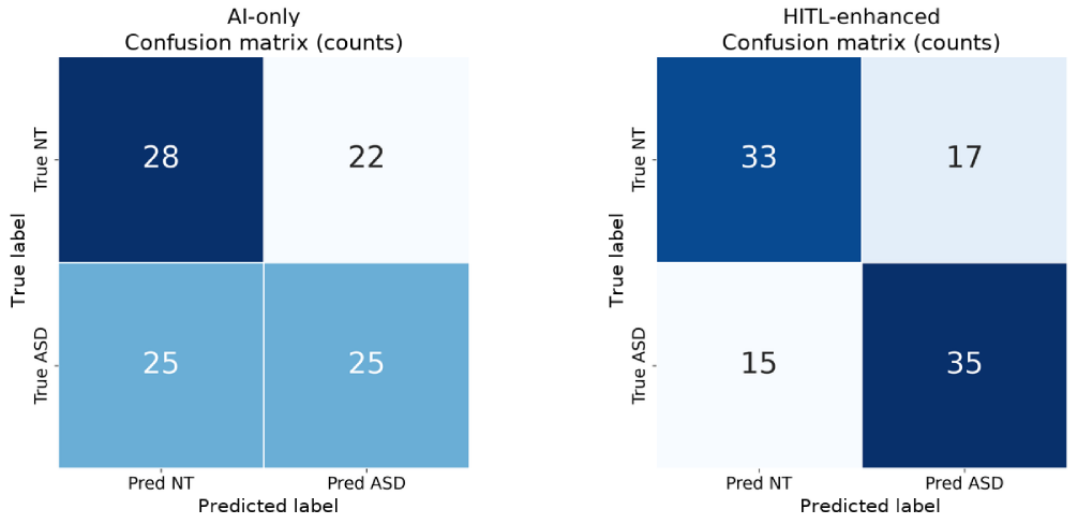
Statistical Validation and Error Analysis

We quantified uncertainty using the aggregate out-of-fold prediction set from the same 5-fold cross-validation procedure. Each fold contained 20 held-out videos, giving 1 held-out prediction for each of the 100 videos. Wilson score 95% CIs were calculated for accuracy, sensitivity, and specificity, while percentile bootstrap 95% CIs were calculated for F_1-score and AUROC. For the HITL-enhanced multimodal model, using humanized FEFE features at a facial confidence threshold of 0.9 together with humanized gaze features, AUROC was 0.76 (95% CI 0.66-0.85), compared with 0.54 (95% CI 0.42-0.65) for the AI-only multimodal model (Figure 8). The intervals displayed for AUROC do not overlap, which is consistent with a higher HITL AUROC; however, CI nonoverlap is not itself a formal hypothesis test.

The aggregate confusion matrix (Figure 9) for the AI-only model contained 28 correctly classified neurotypical videos and 22 neurotypical-to-ASD errors, together with 25 correctly classified ASD videos and 25 ASD-to-neurotypical errors. The 25/25 ASD row represents the model's aggregate performance on 50 ASD videos: sensitivity was 50%, indicating chance-level separation for that class. This is a valid confusion-matrix outcome, not duplicated data. The HITL model correctly classified 33 neurotypical videos and 35 ASD videos, with 17 false positives and 15 false negatives. Total errors decreased from 47 with AI-only features to 32 with HITL features, an approximate 31.9% reduction.

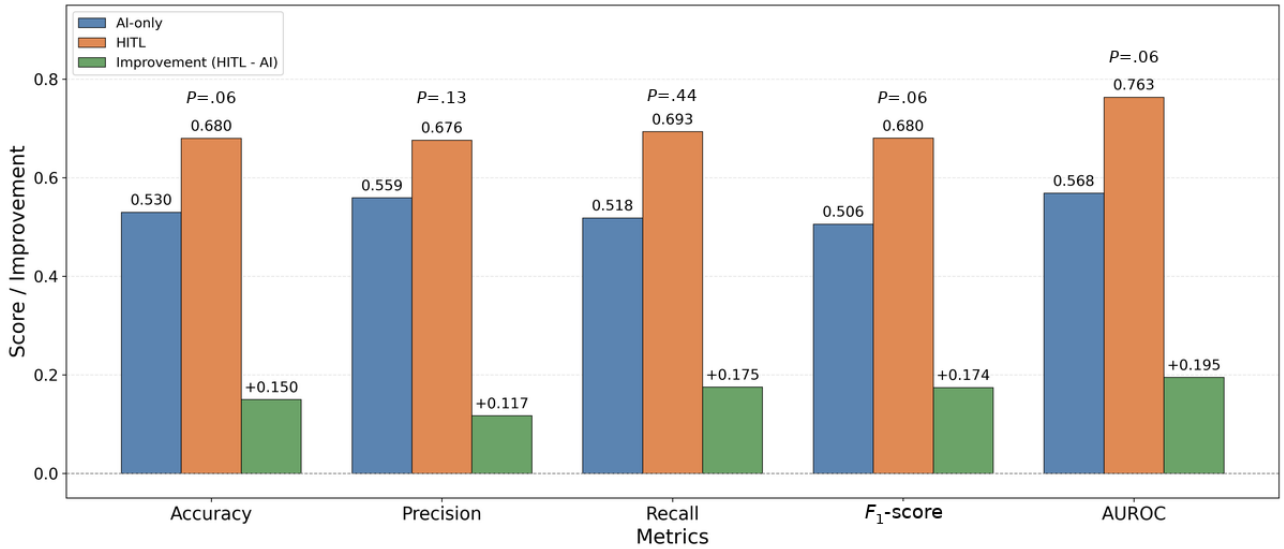
XSL•FO
RenderX

Figure 9. Aggregate confusion matrices for the AI-only multimodal model and the human-in-the-loop–enhanced multimodal model using humanized facial emotion feature extraction at facial confidence threshold of 0.9 and humanized gaze features. ASD: autism spectrum disorder; HILT: human-in-the-loop; NT: neurotypical.



To compare the matched fold-level metrics, we applied 2-sided paired exact Wilcoxon signed-rank tests across the 5 folds (Figure 10). Fold-level comparisons summarize variation across the 5 resampling folds rather than 5 independent experiments. The mean fold-level gains were +0.150 for accuracy, +0.117 for precision, +0.175 for recall, +0.174 for F_1-score, and +0.195 for AUROC. The corresponding P values were .0625 for accuracy, F_1-score, and AUROC, 0.1250 for precision, and 0.4375 for recall. Because only 5 paired folds were available, the minimum attainable exact 2-sided P value was .0625. None of these comparisons reached $P<.05$; therefore, the results should be considered as descriptive fold-level improvements rather than statistically significant effects.

Figure 10. Two-sided paired exact Wilcoxon signed-rank tests comparing AI-only and human-in-the-loop–enhanced multimodal performance across 5 matched cross-validation folds; the minimum attainable P value is .0625. AUROC: area under the receiver operating characteristic curve; HILT: human-in-the-loop.



Discussion

Principal Findings

We present a HITL AI framework for classifying autism from home videos. In our pipeline, AI models extracted behavioral features for each frame of a video, and these AI-extracted features were used for final classification. We replaced these AI-extracted features with human annotations provided by crowd workers in cases where the confidence score of the AI model was below a certain threshold for a particular frame, and we explored the effects of varying the confidence threshold on model performance and the number of human annotations required. We found that, in general, replacing the AI features with more human features tended to improve performance.

Limitations

Several limitations must be acknowledged. First, variability in nonexpert crowd annotations may have introduced noise, despite rigorous selection of crowd workers. Because some workers selected multiple emotion classes and ties were resolved according to the order of the features (an arbitrary process), the

XSL•FO
RenderX

procedure may have been sensitive to column ordering in case of ties. Second, the gaze pipeline used for our secondary analysis did not include a separate NPF or missing-data category like the FEFE analysis did. When Face Mesh did not detect facial landmarks, the implementation returned NaN, representing missing numeric data rather than eye contact or no eye contact. Third, the dataset was limited in size (n=100) despite demographic diversity, which may affect generalizability. Fourth, the binary classification framework (autism vs neurotypical) may have oversimplified autism's heterogeneity; dimensional approaches capturing symptom severity could provide richer phenotypes. Fifth, the temporal modeling capacity of our bidirectional LSTM was constrained by the frame-sampling rate (approximately 0.33 Hz), which was insufficient to capture millisecond-scale dynamics like microexpressions. The model thus likely learned longer-duration behavioral state transitions rather than fine-grained temporal patterns. Sixth, our reliance on 7 basic emotion categories, rather than more granular facial AUs, may have limited sensitivity to subtle, autism-specific differences in facial expressivity [32,34,35]. Emotion labels represent a higher-level abstraction that may not capture the full spectrum of atypical social communication. Seventh, our manual filtration for high video quality (clear facial visibility, center-focused framing) may have introduced selection bias. While necessary for reliable AI-based feature extraction, this process could have systematically excluded data from children with more pronounced symptoms, such as those who frequently avoid the camera or exhibit high motor restlessness. This may limit the tool's applicability to the full clinical spectrum of autism. Eighth, the construct validity of using a game-based task to measure social affect warrants consideration. In the charades gameplay, children were prompted to enact specific emotions. Consequently, the FEFE model assessed the child's ability to comply with instructions and produce recognizable posed expressions, rather than purely spontaneous social affect. Thus, our trained models do not necessarily generalize to naturalistic videos that are not constrained by specific gameplay instructions. Ninth, average precision peaked at a threshold of 0.8 (0.73) and was slightly lower at a threshold of 0.9 (0.72). This indicates that threshold selection involved a trade-off across performance measures and should be evaluated in an independent held-out test set in future work.

Finally, we must note that our preliminary experiments with 4 additional behavioral models yielded inconclusive results, highlighting key challenges in multimodal phenotyping from unstructured home videos. First, the Behavioral Activity Detection Model using ResNet-3D pretrained on Kinetics-400 [36,37] failed to align with our data; selected activity classes (eg, applauding and laughing) often mismatched with extracted frames, which frequently showed unrelated or ambiguous movements. Second, the Audio-Linguistic Recognition Model using Wav2Vec2-Base-960h [38] and Emotion-English-DistilRoBERTa [39,40] could not reliably extract emotional content, as audio segments lacked distinct or consistent vocal expressions matching the predefined emotion classes. Third, the Gesture-Centric Behavioral Model using ResNet-18 pretrained [41] on ImageNet failed to accurately classify gestures such as pointing or waving, as many frames lacked explicit depictions or showed movements too subtle for reliable detection. Fourth, a Self-Stimulatory Behavior (SSB) Model we trained using MobileNetV2 [42] could not detect target behaviors (headbanging, spinning, and arm-flapping) due to their rarity or ambiguity in the sampled frames. These series of negative results on other AI-related features that we explored highlights ongoing challenges with leveraging domain-specific AI models to extract features from video. We note that newer vision language models (VLMs), which are like large language models (LLMs) but take a video or image as an input (in addition to a prompt), can potentially solve these issues and should certainly be explored in future work.

Future Directions

There are several opportunities for future work. First, expanding the dataset with a broader spectrum of participants and incorporating multimodal data (eg, audio, gestures, and wearable sensors) will enhance robustness and generalizability [43]. Second, using expert annotations or multiple annotators per frame would enable formal interrater reliability assessment and improve annotation quality. Third, developing autism-specific models rather than using generic pretrained models could improve base AI performance and reduce the human annotation burden, particularly for behavioral activity, audio-linguistic, gesture, and SSB detection. Fourth, incorporating AU-based features would provide more granular facial movement analysis and potentially greater sensitivity to atypical expressivity than emotion feature vectors. Fifth, future implementations should seek to develop more robust models capable of handling variable video quality to capture behavioral phenotypes in less constrained settings, addressing the selection bias introduced by our quality filtration approach. Sixth, longitudinal studies are needed to evaluate the clinical scalability and real-world effectiveness of HITL-enhanced detection over time. Seventh, future studies should directly compare phenotypes extracted from posed (ie, the charades data collection game) versus spontaneous (ie, the quizzes data collection game) video segments within the same dataset to better isolate markers of spontaneous social engagement specific to autism. Eighth, optimizing the human-AI cost-accuracy balance through systematic cost-benefit analyses will be essential for real-world deployment, particularly given the substantial level of valid human FEFE replacement at the highest observed-performing threshold.

Conclusions

In this proof-of-concept study, the evaluated confidence-driven HITL framework was associated with higher observed autism classification performance in this dataset when a higher proportion of AI features were replaced with human features. That is, among 9 evaluated facial confidence thresholds, the threshold of 0.9 had the highest observed mean F_1-score and AUROC and a valid human FEFE replacement ratio of 59.7%. While these exploratory findings describe our novel HITL AI approach, they do not establish the independent causal effect of human replacement, clinical diagnostic validity, or generalizability to less constrained recordings.

XSL•FO
RenderX

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We acknowledge the use of AI tools (ChatGPT) to edit the language of the manuscript; we emphasize that all ideas and analyses are original and not AI-generated. Figure S1 in Multimedia Appendix 1 was created using the generative AI tool DALL-E 3 by OpenAI.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive nature of the data (ie, pediatric population and protected health information) but are available from the corresponding author upon reasonable request.

Funding

This work was funded by the National Institutes of Health (DP2EB035858, R01LM014342, R01LM013364) and the Stanford Center for Digital Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

Conceptualization: PW, DW
Data curation: KD, AH, AK, DW
Formal analysis: DW, PW, AR
Funding acquisition: PW
Investigation: PW, DW
Methodology: AR, PW
Project administration: PW
Resources: PW, AR
Supervision: PW
Validation: PW, AR
Visualization: AR
Writing – original draft: AR
Writing – review & editing: AR, PW, DW, PA

Conflicts of Interest

None declared.

Multimedia Appendix 1

Deatils of the eye-gaze and facial emotion models, human-in-the-loop replacement workflows, crowd-worker evaluations, threshold performance sweeps, precision-recall curves, and pure-human baseline evaluations.
[DOCX File , 5771 KB-Multimedia Appendix 1]

References

1. Perochon S, Di Martino JM, Carpenter KLH, Compton S, Davis N, Eichner B, et al. Early detection of autism using digital behavioral phenotyping. Nat Med. 2023;29(10):2489-2497. [FREE Full text] [doi: 10.1038/s41591-023-02574-3] [Medline: 37783967]
2. Daniels J, Schwartz JN, Voss C, Haber N, Fazel A, Kline A, et al. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. NPJ Digit Med. 2018;1:32. [FREE Full text] [doi: 10.1038/s41746-018-0035-3] [Medline: 31304314]
3. Hyman SL, Levy SE, Myers SM, COUNCIL ON CHILDREN WITH DISABILITIES, SECTION ON DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics. 2020;145(1):e20193447. [doi: 10.1542/peds.2019-3447] [Medline: 31843864]
4. Shaw KA, Williams S, Patrick ME, Valencia-Prado M, Durkin MS, Howerton EM, et al. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years - autism and developmental disabilities monitoring network, 16 sites, United States, 2022. MMWR Surveill Summ. 2025;74(2):1-22. [doi: 10.15585/mmwr.ss7402a1] [Medline: 40232988]
5. Ning M, Daniels J, Schwartz J, Dunlap K, Washington P, Kalantarian H, et al. Identification and quantification of gaps in access to autism resources in the United States: an infodemiological study. J Med Internet Res. 2019;21(7):e13094. [FREE Full text] [doi: 10.2196/13094] [Medline: 31293243]
6. Geschwind DH. Advances in autism. Annu Rev Med. 2009;60:367-380. [FREE Full text] [doi: 10.1146/annurev.med.60.053107.121225] [Medline: 19630577]

XSL•FO
RenderX

7. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. JAMA Pediatr. 2019;173(5):446-454. [FREE Full text] [doi: 10.1001/jamapediatrics.2019.0285] [Medline: 30907929]
8. Washington P, Kalantarian H, Tariq Q, Schwartz J, Dunlap K, Chrisman B, et al. Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. J Med Internet Res. 2019;21(5):e13668. [FREE Full text] [doi: 10.2196/13668] [Medline: 31124463]
9. Lai MC, Lombardo MV, Auyeung B, Chakrabarti B, Baron-Cohen S. Sex/gender differences and autism: setting the scene for future research. J Am Acad Child Adolesc Psychiatry. 2015;54(1):11-24. [FREE Full text] [doi: 10.1016/j.jaac.2014.10.003] [Medline: 25524786]
10. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall DP. Guess What?: towards understanding autism from structured video using facial affect. J Healthc Inform Res. 2019;3(1):43-66. [FREE Full text] [doi: 10.1007/s41666-018-0034-9] [Medline: 33313475]
11. Pierce K, Gazestani VH, Bacon E, Barnes CC, Cha D, Nalabolu S, et al. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. JAMA Pediatr. 2019;173(6):578-587. [FREE Full text] [doi: 10.1001/jamapediatrics.2019.0624] [Medline: 31034004]
12. Elbattah M, Guérin JL, Carette R, Cilia F, Dequen G. NLP-based approach to detect autism spectrum disorder in saccadic eye movement. IEEE; 2020. Presented at: IEEE Symposium Series on Computational Intelligence (SSCI); December 01-04, 2020:1581-1587; Canberra, ACT, Australia. [doi: 10.1109/ssci47803.2020.9308238]
13. Washington P, Wall DP. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. Annu Rev Biomed Data Sci. 2023;6:211-228. [FREE Full text] [doi: 10.1146/annurev-biodatasci-020722-125454] [Medline: 37137169]
14. Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. Int J Med Inform. 2019;129:29-36. [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.006] [Medline: 31445269]
15. Awaji B, Senan EM, Olayah F, Alshari EA, Alsulami M, Abosaq HA, et al. Hybrid techniques of facial feature image analysis for early detection of autism spectrum disorder based on combined CNN features. Diagnostics (Basel). 2023;13(18):2948. [FREE Full text] [doi: 10.3390/diagnostics13182948] [Medline: 37761315]
16. Wang R, Chaudhari P, Davatzikos C. Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies. Proc Natl Acad Sci U S A. 2023;120(6):e2211613120. [FREE Full text] [doi: 10.1073/pnas.2211613120] [Medline: 36716365]
17. Lakhan A, Mohammed MA, Abdulkareem KH, Hamouda H, Alyahya S. Autism spectrum disorder detection framework for children based on federated learning integrated CNN-LSTM. Comput Biol Med. 2023;166:107539. [doi: 10.1016/j.compbiomed.2023.107539] [Medline: 37804778]
18. Jaiswal A, Kruiper R, Rasool A, Nandkeolyar A, Wall DP, Washington P. Digitally diagnosing multiple developmental delays using crowdsourcing fused with machine learning: protocol for a human-in-the-loop machine learning study. JMIR Res Protoc. 2024;13:e52205. [FREE Full text] [doi: 10.2196/52205] [Medline: 38329783]
19. Washington P. A perspective on crowdsourcing and human-in-the-loop workflows in precision health. J Med Internet Res. 2024;26:e51138. [FREE Full text] [doi: 10.2196/51138] [Medline: 38602750]
20. Myers E, Stone WL, Bernier R, Lendvay T, Comstock B, Cowan C. The diagnosis conundrum: comparison of crowdsourced and expert assessments of toddlers with high and low risk of autism spectrum disorder. Autism Res. 2018;11(12):1629-1634. [doi: 10.1002/aur.2030] [Medline: 30475456]
21. Barai P, Leroy G, Bisht P, Rothman JM, Lee S, Andrews J, et al. Crowdsourcing with enhanced data quality assurance: an efficient approach to mitigate resource scarcity challenges in training large language models for healthcare. AMIA Jt Summits Transl Sci Proc. 2024;2024:75-84. [FREE Full text] [Medline: 38827063]
22. Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, et al. Labeling images with facial emotion and the potential for pediatric healthcare. Artif Intell Med. 2019;98:77-86. [FREE Full text] [doi: 10.1016/j.artmed.2019.06.004] [Medline: 31521254]
23. Raj S, Masood S. Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science. 2020;167:994-1004. [FREE Full text] [doi: 10.1016/j.procs.2020.03.399]
24. Grossard C, Dapogny A, Cohen D, Bernheim S, Juillet E, Hamel F, et al. Children with autism spectrum disorder produce more ambiguous and less socially meaningful facial expressions: an experimental study using random forest classifiers. Mol Autism. 2020;11(1):5. [FREE Full text] [doi: 10.1186/s13229-020-0312-2] [Medline: 31956394]
25. Shi ZH, Groechel TR, Jain S, Chima K, Rudovic OO, Matarić MJ. Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism. J. Hum.-Robot Interact. 2022;11(4):1-28. [doi: 10.1145/3526111]
26. Chi NA, Washington P, Kline A, Husic A, Hou C, He C, et al. Classifying autism from crowdsourced semistructured speech recordings: machine learning model comparison study. JMIR Pediatr Parent. 2022;5(2):e35406. [FREE Full text] [doi: 10.2196/35406] [Medline: 35436234]

XSL•FO
RenderX

27. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, et al. Improved digital therapy for developmental pediatrics using domain-specific artificial intelligence: machine learning study. JMIR Pediatr Parent. 2022;5(2):e26760. [FREE Full text] [doi: 10.2196/26760] [Medline: 35394438]
28. Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. Lancet. 2018;392(10146):508-520. [FREE Full text] [doi: 10.1016/S0140-6736(18)31129-2] [Medline: 30078460]
29. justinshenk/fer: Zenodo. Zenodo. 2021. URL: https://doi.org/10.5281/zenodo.5362356 [accessed 2026-09-11]
30. Goodfellow IJ, Erhan D, Luc Carrier P, Courville A, Mirza M, Hamner B, et al. Challenges in representation learning: a report on three machine learning contests. Neural Netw. 2015;64:59-63. [doi: 10.1016/j.neunet.2014.09.005] [Medline: 25613956]
31. Ekman P. Facial expression and emotion. Am Psychol. 1993;48(4):384-392. [doi: 10.1037//0003-066x.48.4.384] [Medline: 8512154]
32. Du M, Shi P, Liu Z, Xu Y, Liu X, Liu W, et al. Naturalistic facial dynamics enable quantitative clinical assessment of atypical expression phenotypes in children with autism spectrum disorder. NPJ Digit Med. 2026;9(1):183. [FREE Full text] [doi: 10.1038/s41746-026-02375-1] [Medline: 41559146]
33. Banerjee A, Mutlu OC, Kline A, Surabhi S, Washington P, Wall DP. Training and profiling a pediatric facial expression classifier for children on mobile devices: machine learning study. JMIR Form Res. 2023;7:e39917. [FREE Full text] [doi: 10.2196/39917] [Medline: 35962462]
34. Guha T, Yang Z, Grossman RB, Narayanan SS. A computational study of expressive facial dynamics in children with autism. IEEE Trans Affect Comput. 2018;9(1):14-20. [FREE Full text] [doi: 10.1109/TAFFC.2016.2578316] [Medline: 29963280]
35. Ji Y. Hugging rain man: a novel facial action units dataset for analyzing atypical facial expressions in children with autism spectrum disorder. IEEE Transactions on Affective Computing. 2025;16(03):2287-2302. [doi: 10.1109/taffc.2025.3558914/mm1]
36. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. IEEE; 2017. Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017:21-26; Honolulu, HI, USA. [doi: 10.1109/CVPR.2017.502]
37. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. 2017. Presented at: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW); October 22-29, 2017:3154-3160; Venice, Italy. [doi: 10.1109/iccvw.2017.373]
38. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. 2020. Presented at: NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 6-12, 2020:12449-12460; Vancouver BC Canada. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf
39. Emotion English distilRoBERTa-base. Hugging Face. URL: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ [accessed 2026-08-22]
40. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019. [doi: 10.5260/chara.21.2.8]
41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016. Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016:770-778; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.90]
42. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. 2018. Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018:4510-4520; Salt Lake City, UT, USA. [doi: 10.1109/CVPR.2018.00474]
43. Pang L, Zhao X, Zhao L, Li J, Kuo F, Wang H, et al. Multi-modal data analysis for autism spectrum disorder in children: state of the art and trends. EngMedicine. 2026;3(1):100117. [FREE Full text] [doi: 10.1016/j.engmed.2025.100117]

Abbreviations

ADI-R: Autism Diagnostic Interview–Revised
ADOS-2: Autism Diagnostic Observation Schedule, Second Edition
ASD: autism spectrum disorder
AU: action unit
AUROC: area under the receiver operating characteristic curve
EG: eye gaze
FEFE: facial emotion feature extraction
FER: facial expression recognition
HITL: human-in-the-loop
LLM: large language model
LSTM: long short-term memory

XSL•FO
RenderX

NPF: no or partial face
SSB: self-stimulatory behavior
VLM: vision language model

Edited by A Coristine; submitted 29.Sep.2025; peer-reviewed by B Bao, R Kurmashev; comments to author 15.Jan.2026; revised version received 03.Sep.2026; accepted 09.Sep.2026; published 20.Nov.2026

Please cite as:
Rasool A, Dunlap K, Husic A, Kline A, Azizian P, Wall DP, Washington P
Balancing Human and AI Features for Digital Autism Diagnostics from Home Videos: Human-In-The-Loop Machine Learning Study
JMIR AI 2026;5:e84699
URL: https://ai.jmir.org/2026/1/e84699
doi: 10.2196/84699
PMID:

©Abdur Rasool, Kaitlyn Dunlap, Arman Husic, Aaron Kline, Parnian Azizian, Dennis P Wall, Peter Washington. Originally published in JMIR AI (https://ai.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX