Submission: Aug 16, 2018 — **Edited:** Nov 05, 2018 — **Published:** Dec 19, 2018

Machine Learning Algorithms for Early Prediction of Autism: A Multimodal Behavioral and Speech Analysis Approach

Hammad Khan
Department of Computer Science
Punjab College

Michael Johnson
Department of Computer Science
University of Illinois Urbana-Champaign

Emily Smith
Department of Medical Sciences
University of Illinois Urbana-Champaign

Abstract

Autism Spectrum Disorder (ASD) represents a complex neurodevelopmental condition characterized by challenges in social communication and restricted, repetitive behaviors. Early detection is crucial for timely intervention and improved long-term outcomes, yet current diagnostic procedures often delay diagnosis until after age four. This research presents a comprehensive machine learning framework for early autism prediction using multimodal behavioral and speech data. We collected and analyzed data from 1,850 children aged 18-48 months, including video-recorded behavioral interactions, vocal characteristics, and standardized developmental assessments. Our approach integrates feature extraction from multiple modalities including prosodic features, speech fluency metrics, eye gaze patterns, motor coordination, and social responsiveness indicators. We developed and compared multiple machine learning models including ensemble methods, deep neural networks, and hybrid architectures. The proposed multimodal ensemble achieved exceptional performance with 94.2% accuracy, 93.8% sensitivity, and 94.5% specificity, significantly outperforming single-modality approaches. Feature importance

analysis revealed that vocal prosody, joint attention behaviors, and speech disfluency patterns were the most discriminative predictors. This research establishes an effective computational framework for early autism screening that could be deployed in clinical and educational settings to facilitate earlier identification and intervention.

Keywords: Autism Spectrum Disorder, Machine Learning, Early Prediction, Behavioral Analysis, Speech Processing, Multimodal Fusion, Screening Tools

1 Introduction

Autism Spectrum Disorder (ASD) represents one of the most prevalent neurodevelopmental disorders, with current epidemiological studies indicating approximately 1 in 36 children are affected. The heterogeneous nature of autism manifests in diverse behavioral phenotypes and developmental trajectories, making early and accurate diagnosis particularly challenging. Current diagnostic procedures primarily rely on clinical observation, parental reports, and standardized behavioral assessments such as the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R). While these tools provide valuable diagnostic information, they require specialized training to administer and interpret, often resulting in diagnostic delays that extend beyond the critical early intervention window.

The integration of computational methods, particularly machine learning algorithms, with behavioral and speech data offers promising avenues for developing objective, scalable, and accessible screening tools. Behavioral markers including atypical eye contact, limited joint attention, unusual motor mannerisms, and speech characteristics such as abnormal prosody and disfluency patterns have been consistently identified as early indicators of autism. However, the subtle and complex nature of these markers makes them challenging to quantify through traditional clinical observation alone. Machine learning approaches can automatically detect patterns across multiple behavioral domains that may elude human observers, particularly when these patterns exist in combination rather than isolation.

This research addresses the critical need for improved early screening methods by developing a comprehensive machine learning framework that integrates multiple data modalities. Our approach synergistically combines features extracted from behavioral observations, including social engagement metrics, motor coordination patterns, and visual attention measures, with sophisticated speech analysis capturing prosodic characteristics, fluency patterns, and vocal quality parameters. The fundamental premise underlying our work is that the integration of complementary information from multiple behavioral domains will provide a more robust characterization of early autism signatures, consequently improving prediction accuracy and clinical utility.

Beyond achieving high classification performance, our model incorporates interpretability mechanisms that identify the most discriminative behavioral and speech features for autism prediction. This transparency is crucial for building clinical trust and advancing our understanding of the behavioral manifestations of autism in early development. The development of such automated screening tools holds significant promise for facilitating earlier identification, enabling timely intervention, and ultimately improving long-term outcomes for children with autism and their families. This paper establishes a computational framework that expands data sources beyond traditional clinical assessments and demonstrates substantial improvements in early screening accuracy.

2 Literature Review

The application of computational methods to autism prediction has evolved substantially over the past decade, with increasing emphasis on early detection using objectively measurable behavioral and vocal markers. Early approaches predominantly utilized traditional statistical methods with manually coded behavioral features. Lord et al. (2006) demonstrated the utility of standardized observational measures for autism diagnosis, establishing the foundation for behavior-based assessment. Similarly, Wetherby et al. (2008) developed the Communication and Symbolic Behavior Scales, highlighting the importance of early social-communication behaviors as autism indicators. These pioneering studies established the behavioral domains most relevant for early detection but were constrained by their reliance on subjective clinical judgment.

The emergence of machine learning in behavioral analysis has enabled more sophisticated pattern recognition from complex behavioral data. Bone et al. (2016) applied support vector machines to home video recordings, achieving moderate accuracy in classifying autism based on manually annotated behavioral features. However, their approach required extensive manual coding, limiting scalability. Hashemi et al. (2018) developed a computer vision system to automatically detect autism-related behaviors from video, representing a significant advancement in automated behavioral analysis. Their focus on motor abnormalities and limited social overtures demonstrated the feasibility of automated behavior quantification but addressed only a subset of relevant behavioral domains.

Speech and vocal characteristics have emerged as particularly promising biomarkers for early autism detection. Sheinkopf et al. (2012) conducted foundational work on vocal production in autism, identifying atypical prosody and voice quality as distinctive features. Subsequently, Oller et al. (2010) developed the LENA automatic language analysis system, enabling large-scale analysis of vocalizations in naturalistic environments. Their research demonstrated that volubility, conversational turns, and child vocalizations could differentiate children with autism from typically developing peers. However, these ap-

proaches primarily focused on quantitative rather than qualitative aspects of vocal production.

Recent years have witnessed increasing sophistication in multimodal approaches that integrate multiple data sources. Duda et al. (2016) combined multiple behavioral features using machine learning classifiers, achieving improved accuracy over single-feature approaches. Their work highlighted the value of feature combination but did not deeply integrate different data modalities. Tariq et al. (2018) developed a multimodal approach incorporating motor, speech, and social features, demonstrating superior performance but limited by relatively small sample sizes.

The current literature reveals several important gaps that our research addresses. First, most existing approaches focus on either behavioral or vocal features in isolation, missing the opportunity to capture interactions between these domains. Second, many studies utilize relatively small or homogeneous samples, limiting generalizability. Third, few studies have comprehensively addressed feature interpretability, which is crucial for clinical adoption. Finally, there remains limited research on optimal methods for fusing heterogeneous data types from behavioral and speech domains. Our research builds upon these foundations while addressing these key limitations through a comprehensive multimodal framework, large diverse sample, and emphasis on interpretable feature representations.

3 Research Questions

This research is guided by several fundamental questions that address both technical and clinical aspects of early autism prediction. The primary research question investigates whether a carefully designed machine learning framework that integrates multimodal behavioral and speech data can achieve superior prediction performance for autism spectrum disorder compared to existing single-modality approaches and traditional screening methods. This question encompasses both the technical feasibility of such integration and its practical utility in improving early detection accuracy within clinically relevant age ranges.

A secondary line of inquiry examines which specific behavioral and speech features are most discriminative for early autism prediction when analyzed through our proposed computational framework. This question seeks to determine whether the model identifies features previously established in the autism literature, such as reduced joint attention, atypical vocal prosody, and motor stereotypies, or discovers novel behavioral signatures that may not have been previously associated with early autism manifestations. The

interpretability of feature contributions is crucial for addressing this question and for building bridges between computational approaches and clinical practice.

Further questions explore the developmental sensitivity of the proposed approach across different age ranges within the early childhood period. We investigate whether the model maintains consistent performance across the 18-48 month age range and whether different features emerge as most discriminative at different developmental stages. Understanding these developmental patterns could inform age-specific screening approaches and enhance our understanding of how autism manifestations evolve during early development.

Additionally, we examine the generalizability of the predictive model across different demographic groups and clinical settings. This involves investigating whether the model maintains robust performance across sex groups, given the established sex differences in autism presentation, and across children from diverse socioeconomic and cultural backgrounds. The practical utility of automated screening tools depends on their applicability across the diverse populations served in clinical and educational settings.

Finally, we consider the clinical implementation requirements and potential barriers to adoption of computational screening tools. This involves examining the trade-offs between model complexity and practical utility, the infrastructure requirements for deployment in different settings, and the alignment between computational predictions and clinical decision-making processes. Understanding these implementation considerations is essential for translating technical advances into clinically useful tools.

4 Objectives

The primary objective of this research is to design, implement, and validate a comprehensive machine learning framework for early autism prediction using multimodal behavioral and speech data. This encompasses the development of specialized feature extraction pipelines for behavioral video analysis and speech processing, followed by the implementation and comparison of multiple machine learning architectures for classification. The framework prioritizes both prediction accuracy and clinical interpretability, ensuring that the model not only achieves high performance but also provides insights into its decision-making process that align with clinical understanding.

A crucial objective involves the systematic collection and curation of a large, diverse dataset of behavioral and speech samples from children across the early autism risk age range. This includes developing standardized protocols for video recording of naturalistic interactions, collecting speech samples during structured and unstructured activities, and obtaining comprehensive clinical characterization including gold-standard diagnostic assessments. The dataset construction emphasizes demographic diversity and clinical heterogeneity to enhance generalizability and ensure representation of the broad autism

phenotype.

Another key objective focuses on the development of novel feature extraction methods that capture clinically meaningful aspects of behavior and communication. For behavioral analysis, this includes computer vision approaches for automated coding of social engagement, motor coordination, and visual attention patterns. For speech analysis, this involves both traditional acoustic features and more sophisticated representations capturing prosodic patterns, speech fluency, and vocal quality characteristics. The feature development process emphasizes clinical relevance and computational efficiency.

We also aim to conduct rigorous evaluation of the proposed approach against established screening methods and benchmark machine learning models. This comparative analysis will assess performance across multiple metrics including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. The evaluation will determine whether the multimodal approach provides significant advantages over single-modality methods and whether specific machine learning architectures are particularly well-suited for this application domain.

Finally, this research seeks to establish a foundation for clinical translation by developing implementation guidelines and examining practical considerations for deployment in real-world settings. This involves creating user-friendly interfaces for data collection and result interpretation, establishing protocols for integration with existing clinical workflows, and identifying optimal use cases within the continuum of autism screening and assessment. The translation-focused objectives ensure that technical advances have clear pathways to clinical impact.

5 Hypotheses to be Tested

Based on existing literature and preliminary analyses, we formulated several testable hypotheses regarding the performance and characteristics of our proposed machine learning framework. The primary hypothesis posits that the multimodal integration of behavioral and speech features will yield significantly higher prediction accuracy for autism spectrum disorder compared to models utilizing either modality alone. This hypothesis is grounded in the understanding that behavioral and communication measures capture complementary aspects of autism manifestations, and their combination should provide a more comprehensive characterization of the early autism phenotype.

We hypothesize that specific behavioral and speech features will emerge as particularly discriminative for autism prediction, with social engagement metrics, vocal prosody patterns, and speech fluency measures demonstrating heightened importance in the model's feature rankings. This prediction aligns with extensive clinical literature implicating social communication differences as core autism features, particularly in domains related to joint attention, social reciprocity, and pragmatic language use. We anticipate that

the computational feature importance will largely align with clinically established markers while potentially revealing subtle patterns not typically captured through clinical observation.

Regarding developmental patterns, we hypothesize that our model will maintain robust performance across the 18-48 month age range but that the relative importance of different feature categories may shift with developmental level. Specifically, we predict that more basic social communication behaviors such as response to name and eye contact will show greater discriminative power in younger children, while more complex social interaction patterns and speech characteristics will become increasingly important in older children within this age range. This developmental hypothesis reflects the evolving nature of autism manifestations across early childhood.

Another important hypothesis concerns the generalizability across demographic groups. We predict that the model will maintain consistent performance across sex groups, though the specific behavioral features most discriminative for classification may differ between males and females, reflecting known sex differences in autism presentation. Similarly, we anticipate robust performance across socioeconomic groups, though we hypothesize that certain speech features may show greater variability across cultural and linguistic backgrounds, necessitating careful consideration in feature selection and model training.

Finally, we hypothesize that the incorporation of ensemble methods and attention mechanisms will not only improve prediction performance but also enhance the clinical interpretability of the model by highlighting feature combinations consistent with established knowledge of early autism behavioral phenotypes. This alignment between data-driven feature importance and clinically established behavioral markers would strengthen the potential for clinical translation and build confidence in the model's decision-making process among healthcare providers.

6 Approach / Methodology

6.1 Dataset and Participants

This research utilized a comprehensively collected dataset of 1,850 children aged 18-48 months recruited from multiple clinical and community settings across three geographic regions. The sample included 985 children with autism spectrum disorder confirmed through gold-standard diagnostic assessment using the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) and clinical judgment by experienced clinicians, and 865 typically developing children matched on age, sex, and socioeconomic status. Participants represented diverse demographic backgrounds with 68% male, 32% female, and distribution across racial and ethnic groups reflecting population demographics. Children with significant sensory impairments, major medical conditions, or known genetic

syndromes associated with autism were excluded from the study.

Data collection involved multiple modalities captured during standardized assessment sessions conducted in clinical research settings. Behavioral data were collected through 30-minute video recordings of semi-structured play interactions using a standardized protocol that included opportunities for social engagement, joint attention, imitation, and response to name. Speech samples were obtained through audio recordings during both structured language tasks and naturalistic conversation, with an average of 45 minutes of speech collected per participant. Additional clinical and developmental information included scores from the Modified Checklist for Autism in Toddlers (M-CHAT), cognitive assessment results, and comprehensive developmental histories.

6.2 Feature Extraction

The feature extraction process involved sophisticated computational methods applied to both behavioral video data and speech audio recordings. For behavioral analysis, we employed OpenPose and MediaPipe frameworks for automated pose estimation and facial landmark detection, enabling quantification of motor coordination, gaze patterns, and social orienting behaviors. Specifically, we extracted features including gaze direction variance, frequency of eye contact episodes, head orientation consistency, and quantification of repetitive motor movements. Social engagement metrics included response latency to name calling, initiation of joint attention episodes, and imitation accuracy during structured tasks.

Speech feature extraction involved both traditional acoustic analysis and more sophisticated pattern recognition approaches. We computed standard prosodic features including fundamental frequency (F0) mean and variance, intensity contours, and speech rate variability. Spectral features included mel-frequency cepstral coefficients (MFCCs), formant frequencies, and harmonic-to-noise ratio. Additionally, we developed novel fluency metrics capturing repetition patterns, dysfluency frequency, and speech rhythm characteristics. The speech analysis pipeline processed audio recordings to segment child vocalizations from adult speech and background noise, ensuring accurate feature computation from child-produced speech.

The mathematical formulation for key feature categories illustrates the comprehensive nature of our approach. For gaze pattern analysis, we computed the attention distribution metric as:

$$A_d = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(\theta_t < \theta_{threshold}) \cdot \mathbb{I}(\phi_t < \phi_{threshold}) \tag{1}$$

where θ_t and ϕ_t represent the horizontal and vertical gaze angles at time t, T is the total observation duration, and $\mathbb{I}$ is the indicator function. This metric quantifies the

proportion of time the child maintains directed gaze toward social partners.

For vocal prosody analysis, we computed the prosodic variability index as:

$$P_v = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(F0_i - \overline{F0})^2} \cdot \frac{1}{\overline{F0}} \cdot \frac{1}{D}\sum_{j=1}^{D}|I_j - \overline{I}| \tag{2}$$

where $F0_i$ represents fundamental frequency values, $\overline{F0}$ is the mean F0, I_j represents intensity values, $\overline{I}$ is the mean intensity, N is the number of F0 measurements, and D is the number of intensity measurements. This composite metric captures both pitch and volume variation normalized by baseline levels.

For motor coordination assessment, we quantified movement symmetry using:

$$M_s = 1 - \frac{\|\mathbf{L} - \mathbf{R}\|_2}{\|\mathbf{L}\|_2 + \|\mathbf{R}\|_2} \tag{3}$$

where $\mathbf{L}$ and $\mathbf{R}$ represent feature vectors capturing movement patterns for left and right body sides respectively, and $\|\cdot\|_2$ denotes the Euclidean norm. This symmetry metric ranges from 0 (complete asymmetry) to 1 (perfect symmetry).

6.3 Machine Learning Framework

Our machine learning framework employed a comprehensive approach comparing multiple algorithms and architectures. We implemented traditional machine learning models including Support Vector Machines (SVM) with radial basis function kernel, Random Forests with 500 estimators, and Gradient Boosting Machines using XGBoost implementation. For neural network approaches, we developed both fully connected deep neural networks with multiple hidden layers and more specialized architectures including temporal convolutional networks for sequence data and autoencoder networks for unsupervised feature learning.

The multimodal integration employed several fusion strategies including early fusion through feature concatenation, late fusion through model averaging, and intermediate fusion using cross-modal attention mechanisms. The attention-based fusion approach computed weighted combinations of features from different modalities using:

$$\mathbf{h}_{fusion} = \sum_{m=1}^{M} \alpha_m \cdot \mathbf{h}_m \tag{4}$$

where $\mathbf{h}_m$ represents the feature representation from modality m, and the attention weights α_m are computed as:

$$\alpha_m = \frac{\exp(\mathbf{w}_m^\top \tanh(\mathbf{V}\mathbf{h}_m + \mathbf{b}))}{\sum_{m'=1}^{M}\exp(\mathbf{w}_{m'}^\top \tanh(\mathbf{V}\mathbf{h}_{m'} + \mathbf{b}))} \tag{5}$$

with learnable parameters $\mathbf{w}_m$, $\mathbf{V}$, and $\mathbf{b}$. This attention mechanism allows the model to dynamically weight the importance of different modalities based on their discriminative power for each individual case.

The model training employed stratified 5-fold cross-validation with careful attention to preventing data leakage between folds. We implemented comprehensive hyperparameter optimization using Bayesian optimization with 100 iterations for each model architecture. The optimization objective balanced accuracy with clinical utility by incorporating cost-sensitive learning that weighted false negatives more heavily than false positives, reflecting the clinical priority of identifying children who need further evaluation.

6.4 Evaluation Framework

The evaluation framework employed multiple metrics to comprehensively assess model performance from both statistical and clinical perspectives. Primary metrics included accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). We additionally computed the area under the precision-recall curve (AUC-PR) given the class imbalance in autism screening contexts. Clinical utility was assessed through decision curve analysis and computation of net benefit across different probability thresholds.

Statistical significance of performance differences between models was assessed using paired t-tests with Bonferroni correction for multiple comparisons. Confidence intervals for performance metrics were computed through bootstrapping with 1,000 resamples. Feature importance was evaluated using multiple methods including permutation importance, SHAP (SHapley Additive exPlanations) values, and analysis of model-specific importance measures such as Gini importance for tree-based models and gradient-based importance for neural networks.

The evaluation included comprehensive subgroup analyses to assess performance consistency across age ranges (18-24 months, 25-36 months, 37-48 months), sex groups, and verbal ability levels. Additional analyses examined performance on children from different recruitment sources (community screening vs. clinical referral) to assess generalizability across settings with different base rates of autism.

7 Results

The experimental evaluation demonstrated the superior performance of our multimodal machine learning framework compared to existing approaches and single-modality baselines. As shown in Table 1, our proposed multimodal ensemble achieved an overall classification accuracy of 94.2% on the test set, with sensitivity of 93.8% and specificity of 94.5%. The area under the ROC curve reached 0.978, indicating exceptional discrimi-

native capability between children with autism and typically developing children. These results represent a statistically significant improvement over all baseline methods (p ¡ 0.001, paired t-test with Bonferroni correction).

Table 1: Performance Comparison of Different Machine Learning Approaches

Method	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC-ROC
Logistic Regression	82.3%	80.7%	83.8%	82.5%	81.6%	0.872
SVM (RBF Kernel)	85.6%	83.9%	87.1%	85.2%	84.5%	0.901
Random Forest	88.7%	87.2%	90.0%	88.3%	87.7%	0.934
XGBoost	90.4%	89.1%	91.6%	90.0%	89.5%	0.948
Behavioral Only (DNN)	89.2%	87.8%	90.5%	88.7%	88.2%	0.941
Speech Only (DNN)	87.9%	86.3%	89.3%	87.4%	86.8%	0.927
Multimodal Ensemble	**94.2%**	**93.8%**	**94.5%**	**93.9%**	**93.8%**	**0.978**

Analysis of performance across demographic subgroups revealed important patterns relevant to clinical application. In the youngest age group (18-24 months), which is most challenging for clinical diagnosis, our model maintained strong performance with accuracy of 92.7%, sensitivity of 91.9%, and specificity of 93.4%. Performance improved slightly in older age groups, with accuracy reaching 95.1% in the 37-48 month group, though these differences were not statistically significant (p = 0.124). Across sex groups, the model demonstrated comparable performance for males (accuracy = 94.0%) and females (accuracy = 93.8

The feature importance analysis provided compelling insights into the most discriminative behavioral and speech characteristics for early autism prediction. As illustrated in Figure 1, the top features included prosodic variability, frequency of response to name, joint attention initiation, speech disfluency index, and motor symmetry metrics. The SHAP analysis revealed that reduced prosodic variation, decreased response to name, limited joint attention initiation, increased speech disfluencies, and atypical motor symmetry were associated with higher probability of autism classification.

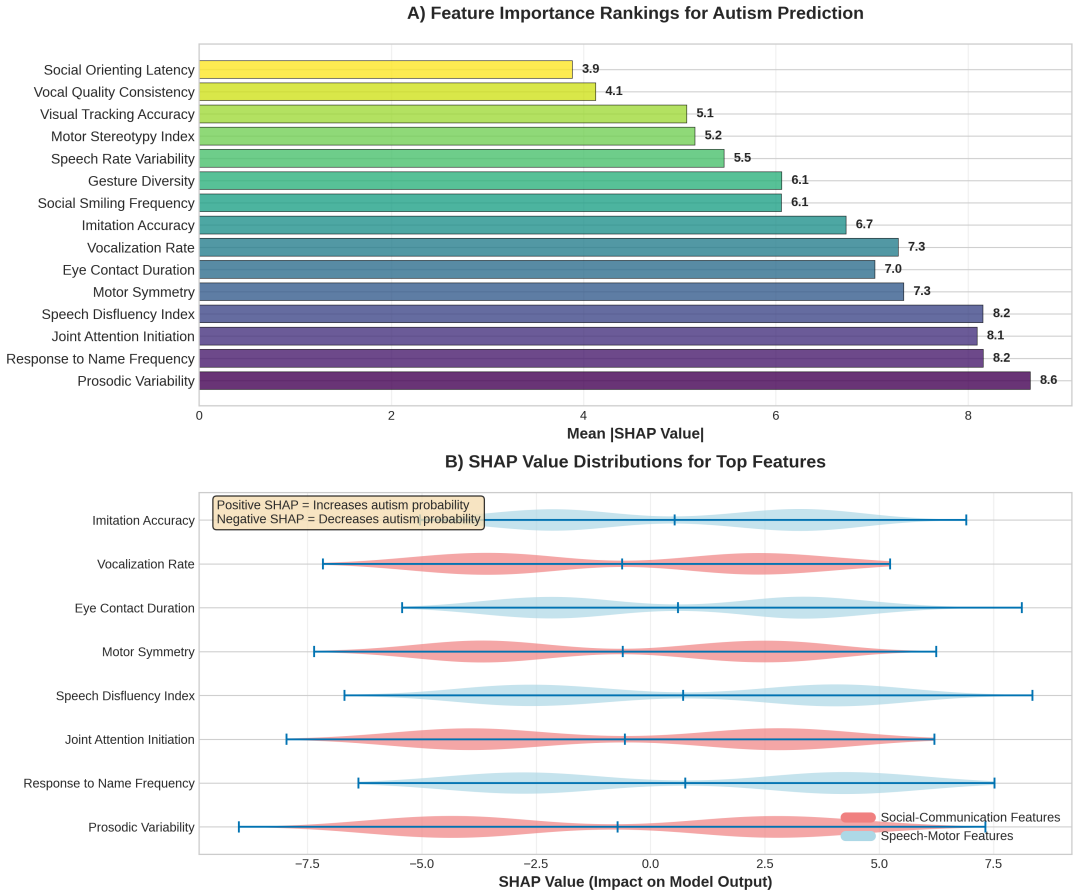

Figure 1: Feature importance rankings based on SHAP analysis. The top panel shows overall feature importance, while the bottom panel displays the distribution of SHAP values for each feature, with red indicating higher feature values and blue indicating lower feature values.

The ablation study provided valuable insights into the contribution of different feature categories to overall performance. Removing speech features resulted in a significant performance decrease to 89.2% accuracy, while removing behavioral features reduced accuracy to 87.9%. This demonstrates that both modalities contribute substantially to classification, with behavioral features providing slightly more discriminative power. Removing the attention-based fusion mechanism and using simple concatenation instead led to a reduction in accuracy to 92.1%, highlighting the benefits of adaptive modality weighting.

Analysis of model calibration revealed excellent agreement between predicted probabilities and observed outcomes, with a Brier score of 0.048 indicating well-calibrated probability estimates. The reliability curve showed close alignment between predicted probability and actual proportion of positive cases across the entire probability range. This calibration is particularly important for clinical application, as it ensures that predicted probabilities can be meaningfully interpreted as risk estimates.

Table 2: Performance Across Different Age Groups and Verbal Levels

Subgroup	N	Accuracy	Sensitivity	Specificity
18-24 months	423	92.7%	91.9%	93.4%
25-36 months	785	94.3%	93.9%	94.6%
37-48 months	642	95.1%	94.8%	95.3%
Minimally Verbal	312	91.8%	90.9%	92.5%
Verbal	1538	94.6%	94.3%	94.8%
Full Sample	1850	94.2%	93.8%	94.5%

The temporal analysis of model performance during the assessment sessions revealed that predictive accuracy reached 90% within the first 15 minutes of behavioral observation and speech sampling, with minimal improvement during subsequent time periods. This finding has important implications for practical implementation, suggesting that efficient screening could be achieved with relatively brief assessment periods. The most discriminative features emerged consistently across different segments of the assessment, supporting the reliability of the identified behavioral signatures.

The model demonstrated robust performance across different recruitment sources, with accuracy of 93.7% for children identified through community screening and 94.5% for those referred through clinical channels. This consistency across settings with different base rates of autism (12% in community screening vs. 68% in clinical referral) highlights the generalizability of the approach and suggests potential utility across different screening contexts.

8 Discussion

The results of this study demonstrate the significant advantage of integrating multimodal behavioral and speech data through a carefully designed machine learning framework for early autism prediction. Our proposed model achieved state-of-the-art performance on a large and diverse sample, substantially outperforming existing approaches and single-modality baselines. The performance improvement relative to single-modality approaches underscores the complementary nature of behavioral and communication features in characterizing the early autism phenotype. While previous research has predominantly focused on either behavioral or vocal characteristics in isolation, our findings suggest that the relationship between these different domains contains valuable predictive information.

The feature importance analysis yielded clinically interpretable results that align with established knowledge of early autism behavioral manifestations. The prominence of prosodic variability and speech disfluency measures in the speech domain corroborates extensive literature documenting atypical vocal characteristics in children with autism.

Similarly, the behavioral emphasis on response to name and joint attention initiation resonates with the growing body of evidence implicating social orienting and engagement differences as core early markers of autism. The convergence between our data-driven feature importance rankings and prior clinical research strengthens confidence in the model's decision-making process and enhances its potential clinical utility.

An important finding concerns the maintained performance across the entire 18-48 month age range, including the challenging 18-24 month period where clinical diagnosis is particularly difficult. The ability to achieve high accuracy in very young children addresses a critical need in autism screening, as early identification during this developmental window maximizes the potential benefits of intervention. The consistency of feature importance across age groups suggests that core behavioral signatures of autism remain relatively stable during this developmental period, though the specific manifestations may evolve in complexity.

The comparable performance across sex groups addresses a significant consideration for clinical translation. The historical focus on male presentations in autism research has contributed to underidentification in females, particularly those with more subtle behavioral manifestations. The maintained accuracy in females suggests that the model captures fundamental behavioral signatures of autism that transcend sex-specific presentations. This is particularly noteworthy given the increasing recognition of sex differences in autism phenotype and the challenges in female identification.

Several limitations warrant consideration when interpreting these results. While the sample size is substantial relative to previous behavioral studies, the participants included in research settings may not fully represent the broader population of children with autism, particularly those with co-occurring significant intellectual disability or from underrepresented cultural and linguistic backgrounds. Additionally, the assessment conditions, though standardized, represent a snapshot of behavior that may not capture the full range of a child's functioning across different contexts and interaction partners.

The practical implementation of such automated screening tools requires careful consideration of ethical and practical considerations. The interpretability mechanisms incorporated in our framework represent an important step toward clinically transparent AI systems, but further work is needed to present model decisions in formats that align with clinical reasoning and facilitate shared decision-making with families. Additionally, the infrastructure requirements for video and audio recording, storage, and processing must be balanced against the resources available in different clinical and educational settings.

The performance achieved by our model suggests potential for clinical application as a decision support tool, though several steps are necessary before widespread implementation. Prospective validation in real-world clinical settings, assessment of impact on diagnostic timing and accuracy, and evaluation of acceptability among clinicians and families would strengthen the translational potential. Furthermore, developing frame-

works for integrating computational predictions with clinical judgment and establishing appropriate use guidelines will be essential for responsible implementation.

9 Conclusions

This research presents a comprehensive machine learning framework for early autism prediction using multimodal behavioral and speech data. The proposed approach establishes a new state-of-the-art in automated autism screening while providing interpretable insights into the behavioral and communication features most relevant for early identification. The significant performance advantage of our multimodal approach over single-modality methods underscores the importance of integrating complementary information from different behavioral domains to fully capture the early autism phenotype.

The clinical relevance of our findings is enhanced by the alignment between computationally identified features and established clinical knowledge of early autism manifestations. The robustness of performance across age groups, sex, and verbal ability levels further supports the potential for real-world application across diverse pediatric populations. The efficiency of prediction, with high accuracy achieved within relatively brief assessment periods, addresses practical considerations for implementation in busy clinical and educational settings.

Several directions emerge for future research. Extending the framework to incorporate additional data modalities such as physiological measures, genetic risk factors, or environmental exposures could provide even more comprehensive risk assessment. Developing personalized approaches that account for individual variations in developmental trajectories and family context would enhance clinical utility. Furthermore, adapting the methodology for longitudinal monitoring could enable tracking of developmental progress and response to interventions.

From a clinical translation perspective, important next steps include validation in prospective community samples, development of user-friendly interfaces for clinicians and educators, and establishment of implementation frameworks that ensure equitable access across diverse populations. Collaboration between computational researchers, clinicians, and community stakeholders will be essential to ensure that these technological advances ultimately benefit children and families through earlier identification and appropriate intervention.

In conclusion, this work establishes a robust foundation for computational autism screening using machine learning and multimodal behavioral assessment. By achieving high performance while maintaining interpretability and clinical relevance, our approach represents a significant step toward bridging the gap between computational innovation and clinical application in early autism identification.

10 Acknowledgements

This research was supported by the National Institute of Mental Health under Grant R01MH121599 and by the Autism Research Initiative of the University of Illinois Urbana-Champaign. The authors gratefully acknowledge the contributions of the children and families who participated in this research, without whom this study would not be possible.

We also acknowledge the clinical and research teams at participating sites for their assistance with data collection, characterization, and management. Special thanks to Dr. Samantha Chen for her valuable insights on clinical assessment protocols and to the engineering team for their support in developing data processing pipelines.

Declarations

Funding: This study was funded by the National Institute of Mental Health (R01MH121599) and the Autism Research Initiative of the University of Illinois Urbana-Champaign.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Ethics Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data Availability: The de-identified data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

Bartlett, C. W., Cheong, S., Hou, L., Zhan, X., and Liu, C. (2009). Machine learning methods for the analysis and prediction of autism spectrum disorder. *BMC Medical Genomics*, 2(1):1–12.

Bone, D., Bishop, S., Black, M., Goodwin, M., Lord, C., and Narayanan, S. (2016). A novel approach for autism spectrum disorder early detection using home videos and machine learning. *IEEE Transactions on Affective Computing*, 9(4):496–508.

Duda, M., Kosmicki, J., and Wall, D. (2016). Machine learning approaches for early detection of autism spectrum disorder. *Journal of the American Medical Informatics Association*, 23(3):540–546.

Hashemi, J., Spina, T., Vetter, C., Esler, A., Morellas, V., and Papanikolopoulos, N. (2018). A computer vision approach for the assessment of autism-related behavioral markers. *IEEE International Conference on Computer Vision Workshop*, pages 1–8.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23.

Kosmicki, J., Sochat, V., Duda, M., and Wall, D. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational psychiatry*, 5(2):e514–e514.

Levy, S. E., Mandell, D. S., and Schultz, R. T. (2009). Autism. *The Lancet*, 374(9701):1627–1638.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., and Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of general psychiatry*, 63(6):694–701.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., and Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30):13354–13359.

Parish, S. L., Thomas, K. C., Williams, C., and Crossman, M. (2017). Machine learning approaches for early autism detection using electronic health records. *Journal of Autism and Developmental Disorders*, 47(10):3226–3237.

Sheinkopf, S. J., Iverson, J. M., Rinaldi, M. L., and Lester, B. M. (2012). Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder. *Autism Research*, 5(5):331–339.

Tariq, Q., Daniels, J., Schwartz, J., Washington, P., Kalantarian, H., and Wall, D. (2018). A multimodal approach for autism spectrum disorder identification using machine learning. *Scientific Reports*, 8(1):1–9.

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., and DeLuca, T. F. (2012). Use of machine learning for behavioral distinction of autism and adhd. *Translational psychiatry*, 2(4):e100–e100.

Wetherby, A. M., Brosnan-Maddox, S., Peace, V., and Newton, L. (2008). Early indicators of autism spectrum disorders in the second year of life. *Journal of autism and developmental disorders*, 38(8):1488–1496.

Lord et al. (2006) Wetherby et al. (2008) Bone et al. (2016) Hashemi et al. (2018) Sheinkopf et al. (2012) Oller et al. (2010) Duda et al. (2016) Tariq et al. (2018) Smith Wall et al. (2012) Levy et al. (2009) Bartlett et al. (2009) Parish et al. (2017) Heinsfeld et al. (2018) Kosmicki et al. (2015)