RESEARCH PAPER

Modelling, Implementation, and Performance Analysis of a Multimodal Expert System for Autism Spectrum Disorder Prediction

Ms. Pournima P Bhangale[1], Dr. Rajendra B Patil[2]
[1]University Department of Information Technology, University of Mumbai, Mumbai, India
Email: ppbhangale@yahoo.com
[2]University Department of Information Technology, University of Mumbai, Mumbai, India
Email: patilrajendrab@gmail.com

ABSTRACT
Autism Spectrum Disorder (ASD) is a complex disorder in neurodevelopment that should be screened early and correctly to facilitate a timely intervention and a better developmental outcome. The conventional diagnostic tests rely mainly on formal behavioural evaluations and specialist clinical opinion, which tend to be both time-intensive, subjective and unavailable. In order to address these constraints, this paper will present a multimodal expert system to predict ASD by combining structured behavioural screening data, unstructured texts reports by caregivers, and demographics as one machine learning system. This system uses naive natural language processing methods and compares various classification models, such as Naive Bayes, Support Vector Machines, Random Forest, XGBoost and deep learning models. Heterogeneous data modalities are combined with the help of feature fusion strategies, and class imbalance is solved by means of resampling and the loss-adjustment. To improve not only the ethical reliability but also the interpretability, fairness auditing and explainable artificial intelligence (XAI) techniques are added. The experimental data on publicly accessible Toddler and Adult ASD datasets show better performance compared to unimodal models, which proves the potential of the framework as a scalable and clinically interpretable decision-support system to detect ASD in the initial stages of its screening.

KEYWORDS: Autism Spectrum Disorder; Multimodal Expert System; Machine Learning; Naïve Natural Language Processing; Explainable AI; Fairness-Aware Prediction

How to cite this article: Bhangale PP, Patil RB. Modelling, Implementation, and Performance Analysis of a Multimodal Expert System for Autism Spectrum Disorder Prediction. Int J Drug Deliv Technol. 2026;16(6s): 944-953; DOI: 10.25258/ijddt.16.6s.123

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by persistent challenges in social interaction, communication, and the presence of restricted or repetitive behaviours. The term "spectrum" reflects the wide variability in symptom severity, functional abilities, and developmental trajectories observed among individuals with ASD. Although global prevalence rates have increased due to improved awareness, broader diagnostic criteria, and enhanced screening practices, early identification remains challenging, particularly in resource-constrained healthcare settings. Early screening is critical because timely intervention during sensitive developmental periods can significantly improve cognitive, social, and adaptive outcomes. Conversely, delayed diagnosis often results in missed intervention opportunities, increased caregiver burden, and long-term economic costs. The early detection of ASD is further complicated by heterogeneous symptom presentation, age-dependent behavioural changes, frequent comorbidities, and cultural influences, underscoring the need for scalable, objective, and data-driven approaches that integrate diverse sources of information for early risk prediction.

Conventional ASD diagnostic methods rely heavily on standardized screening tools, structured interviews, and clinician-based observations. While these instruments demonstrate strong clinical validity, their practical application is often limited by the need for trained specialists, repeated assessments, and prolonged observation periods, which can delay diagnosis. These challenges are particularly pronounced in low- and middle-income regions where access to specialized healthcare is limited. Additionally, traditional diagnostic approaches are inherently subjective, as they depend on caregiver reports and clinical judgment, making them vulnerable to recall bias and cultural variation. Most existing tools also focus on categorical diagnosis rather than continuous risk estimation, reducing their sensitivity to early or borderline cases. Although artificial intelligence approaches have been introduced to address these limitations, early AI-based methods predominantly relied on unimodal data

sources, such as questionnaires, images, speech, or eye-tracking. Such unimodal and often opaque models lack contextual richness, scalability, and interpretability, highlighting the need for transparent multimodal solutions.

The motivation for this study arises from the inherently multimodal nature of ASD prediction. Behavioural screening scores provide quantifiable indicators of risk but fail to capture the linguistic, emotional, and situational nuances present in caregiver narratives and observational reports. Unstructured textual data often contain valuable contextual information regarding communication patterns and behavioural responses that fixed questionnaires cannot fully represent. Demographic metadata further contextualizes behavioural observations and enables fairness-aware evaluation. Multimodal learning offers a principled framework for integrating these heterogeneous data streams, allowing models to exploit complementary information and reduce predictive uncertainty. However, effective clinical deployment requires not only high predictive accuracy but also transparency, structured reasoning, and ethical reliability. This study therefore proposes a systematically modelled multimodal expert system that integrates behavioural, textual, and demographic information within an explainable and fairness-conscious framework. The contributions of this paper include the development of an interpretable multimodal architecture, a comprehensive comparative evaluation of classical and advanced models, and the integration of fairness auditing and explainable AI techniques to support responsible clinical decision-making.

2. Related Work and Background

Autism Spectrum Disorder (ASD) has traditionally been diagnosed using standardized behavioural screening instruments and clinician-led assessments that focus on core symptoms such as impairments in social communication, restricted interests, and repetitive behaviours. Widely adopted tools, including M-CHAT, Q-CHAT, Q-CHAT-10, ADOS, and ADI-R, offer structured and clinically validated frameworks for assessment and diagnosis. Although these instruments demonstrate strong diagnostic reliability, they also suffer from practical limitations. Their administration often requires trained professionals, repeated evaluations, and longitudinal observation, which can delay diagnosis, especially in regions with limited access to specialized healthcare services. In addition, caregiver-reported questionnaires are inherently subjective and influenced by recall bias, cultural context, and individual perceptions of behaviour. Most conventional diagnostic tools emphasize categorical outcomes rather than continuous risk estimation, reducing sensitivity to early or borderline cases. Given the heterogeneous and context-dependent presentation of ASD, these limitations highlight the need for scalable, objective, and data-driven methods that can complement and enhance clinical decision-making.

To address these challenges, machine learning (ML) techniques have been increasingly explored for ASD prediction. Early studies primarily relied on structured behavioural data derived from standardized screening instruments and employed classical classifiers such as Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Trees, and k-Nearest Neighbour, achieving moderate accuracy levels typically between 80% and 88%. Ensemble learning methods, including Random Forest and Gradient Boosting, were later introduced to improve predictive performance by combining multiple learners. Random Forest demonstrated robustness to noisy features, while XGBoost improved generalization through optimized tree construction and regularization. However, these models were largely dependent on structured inputs and failed to capture contextual and linguistic information critical to ASD assessment. More recent deep learning approaches using CNNs, RNNs, and LSTMs on modalities such as images, EEG, and speech have shown promising results but require large datasets, substantial computational resources, and often lack interpretability. The opaque nature of such models raises concerns regarding bias, transparency, and clinical trust.

Natural Language Processing (NLP) has emerged as a valuable approach for ASD prediction by enabling the analysis of unstructured textual data, including caregiver narratives, clinical notes, and interview transcripts. Linguistic features provide insights into communication patterns, emotional expression, and narrative coherence associated with ASD-related behaviours. Early NLP-based studies employed interpretable techniques such as Bag-of-Words and TF-IDF combined with classifiers like Naïve Bayes and SVM. More recent research has explored transformer-based embeddings that capture richer semantic information but often compromise transparency and cross-cultural generalizability. Multimodal ASD detection, which integrates behavioural, textual, demographic, and sensory data, has demonstrated improved robustness and accuracy. However, challenges related to modality alignment, missing data, interpretability, fairness, and real-world deployment remain unresolved. This study addresses these gaps by proposing a fair, explainable multimodal expert system that integrates heterogeneous data sources within an ethically aligned and interpretable decision-support framework.

3. System Modelling and Architecture

3.1 Overview of the Proposed Multimodal Expert System

This proposed system will be a multimodal expert system that will assist in early ASD detection through the introduction of heterogeneous data in a single analytical platform. In contrast to the unimodal models, it uses structured behavioural screening data along with the unstructured caregiver text and demographic metadata to both capture the quantitative indicators and the qualitative diagnostic cues. The use of expert system paradigm facilitates uniform inferences, probabilistic reasoning and comprehensible results to make the framework more of a clinical decision-support tool than an automated diagnostic system. In architecture, the system has four modular layers namely data ingestion and preprocessing, feature extraction and representation, multimodal feature fusion, and expert system inference with explanation production. This is a scalable and modular design that can be expanded to new modalities. The system delivers an ASD risk label which includes confidence scores and descriptions, manages missing or noisy data and meets clinical transparency and accountability demands.

3.2 Data Modalities and Feature Representation

A multimodal expert system is determined by the success of data modality representation and maintenance of complementary information. The proposed system will be integrated with three main modalities, including behavioural and clinical screening capabilities, caregiver report-based text features and demographic metadata. The modalities address very different but connected features of behaviour related to ASD.

3.2.1 Behavioural and Clinical Screening Features

The system has behavioural and clinical screening features as the structured core. Such characteristics are based on standardized ASD screening tools like Q-CHAT-10, M-CHAT and ADOS, which evaluate the key behavioural areas of social engagement, communication skills, attention patterns and repetitive behaviours. The answers are coded in numbers and can be directly incorporated into machine learning.

Along with score scores based on questionnaires, the system accommodates structured behavioural measures based on or derived through clinical observations or video based measures, such as frequency of eye-contact or frequency of gestures, where accessible. Any numerical behavioural characteristics are scaled so that they have uniform scaling across datasets. The system ensures that it is consistent with the current diagnostic practices since it is based on clinically proven screening instruments, and it allows making more predictive modelling due to the integration of data.

3.2.2 Textual Features from Caregiver Reports

Caregiver reports and clinical notes consist of unstructured textual data that contains a lot of contextual information that supplements structured behavioural scores. Caregivers tend to reflect hidden behavioural patterns, feelings, and contexts of situations that might not be reflected in set questionnaires. The system uses naive natural language processing (NLP) techniques in order to include this information; these techniques focus on interpretability and efficiency.

Preprocessing of the texts involve lowercasing, tokenization, removing punctuations, removing stop-words and lemmatization. Textual characteristics are mostly modelled in term frequency-inverse document frequency (TF-IDF) which gives greater weight to discriminative terms across documents. TF-IDF representation is defined to be:

$$TF\text{-}IDF(t,d) = tf(t,d) \cdot \log\left(\frac{N}{df(t)}\right)$$

where $tf(t,d)$ denotes the frequency of term t in document d, $df(t)$ represents the number of documents containing term t, and N is the total number of documents.

Besides TF-IDF vectors, other assistive linguistic features are lexical diversity, sentiment polarity, and simple syntactic cues are extracted. These characteristics can describe the communication patterns that are typically related to ASD, like the lack of diversity of vocabulary or lack of typical emotional displays, and are otherwise clear and understandable.

3.2.3 Demographic and Metadata Features

The demographic and contextual metadata ensure the necessary background data that affected the interpretation of behaviour and enable fairness-conscious assessment. Individual characteristics like age, gender, ethnicity, and economic status are included to put screening results in the context and absorb the variation in development. Of special significance is age since behaviours that are related to ASD have varied differences in all stages of development.

Categorical demographic data are one-hot or label encoded, and continuous data is normalized. Notably, demographic characteristics are never employed to enforce bias but to control and balance predictive performance differences between subgroups. Metadata clearly added facilitates systematic auditing of fairness and alludes that the system is ethical in the clinical decision support.

3.3 Feature Fusion Strategy

One of the major issues of multimodal modelling is the effective combination of heterogeneous features and maintenance of modality-specific information. The offered system is based on the two-tier feature fusion strategy, which balances the simplicity, performance, and interpretability.

3.3.1 Early Fusion (Concatenation)

Early fusion is the standard of the integration approach in which features of all modalities are appended to form one integrated feature set. After the process of normalization, behavioural scores, textual embeddings, demographic metadata, and temporal features are unified:

$$F = [F_b \;||\; F_t \;||\; F_m]$$

where F_b, F_t, and $F_m$represent behavioural, textual, and metadata feature sets, respectively. This approach allows classical machine learning models to directly learn cross-modal relationships within a shared feature space.

Early fusion is highly interpretable and computationally efficient because the contribution of features can be narrowly traced to one of the modalities. It however makes an equal contribution of all modalities and explicitly does not model interactions between modalities. In this way, early fusion is mostly utilized as a robust benchmark on which superior strategies are considered.

3.3.2 Attention-Based Multimodal Fusion

In order to preserve complex interdependence of modalities, the system uses an attention-based multimodal fusion mechanism. The feature representations in this method are represented as modality-specific but then they are mapped into a common latent space, where attention weights are dynamically employed to calculate how much the features matter. The attention operation is formulated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Q, K and V represent query, key and value matrices and d k is the dimension of key vectors.

Attention based fusion allows the system to highlight the informative modalities and de-emphasize the noisy or missing inputs. It is especially helpful in the real-life context where textual reports can be of a different quality or behavioural scores can be incompletely absent. This strategy is more computationally expensive, but is more predictively robust and has interpretable attention weights that can be visualized in order to learn about the contribution of modality.

3.4 Expert System Inference Workflow

The expert system inference process combines the outputs of multimodal models with specific reasoning and explainability processes. After fusion of the feature, trained classifiers generate an ASD risk label and a confidence score. The expert system layer takes these outputs and applies input reasoning rules and decision thresholds to contextualize predictions. The performance validation, fairness auditing and explanation generation are part of the workflow. Measures of fairness are considered based on sensitive demographic characteristics, and intervention mechanisms become triggered when the gaps are too large. Explainability modules have both global and local interpretations, finding influential features and other counterfactual scenarios of interest. The system output final report has three parts namely the projected ASD label, probabilistic confidence score, and a detailed explanation report. In this report, clinicians and other caregivers can get a clear picture of decision drivers, which improves trust and aids informed and ethical decision making.

4. Implementation Details

The proposed multimodal expert system was implemented and evaluated using two publicly available autism spectrum disorder (ASD) screening datasets representing different age groups, namely a Toddler ASD dataset and an Adult ASD dataset. These datasets are widely used in ASD research due to their standardized screening formats and reproducibility. The Toddler dataset comprises over one thousand records derived primarily from Q-CHAT assessments that capture early social interaction, communication behaviours, attention patterns, and repetitive actions. In addition to behavioural indicators, the dataset includes demographic and medical metadata such as age, gender, ethnicity, family history of autism, and neonatal factors. ASD labels are assigned using established screening thresholds, resulting in a binary classification task with moderate class imbalance. The Adult dataset

consists of several hundred cases based on ADOS and M-CHAT assessments, supplemented with demographic attributes and self-reported behavioural characteristics. Compared with the toddler data, the adult dataset exhibits greater behavioural heterogeneity and a skewed distribution of ASD cases. Both datasets were augmented with unstructured caregiver text to support comprehensive multimodal analysis.

To provide a clear architectural understanding of the proposed system, the overall workflow of the multimodal ASD prediction framework is illustrated diagrammatically in **Figure 1**. The framework integrates multimodal data acquisition, preprocessing, feature engineering, multimodal fusion, prediction modelling, explainability analysis, and clinician-assisted decision support within a unified pipeline.

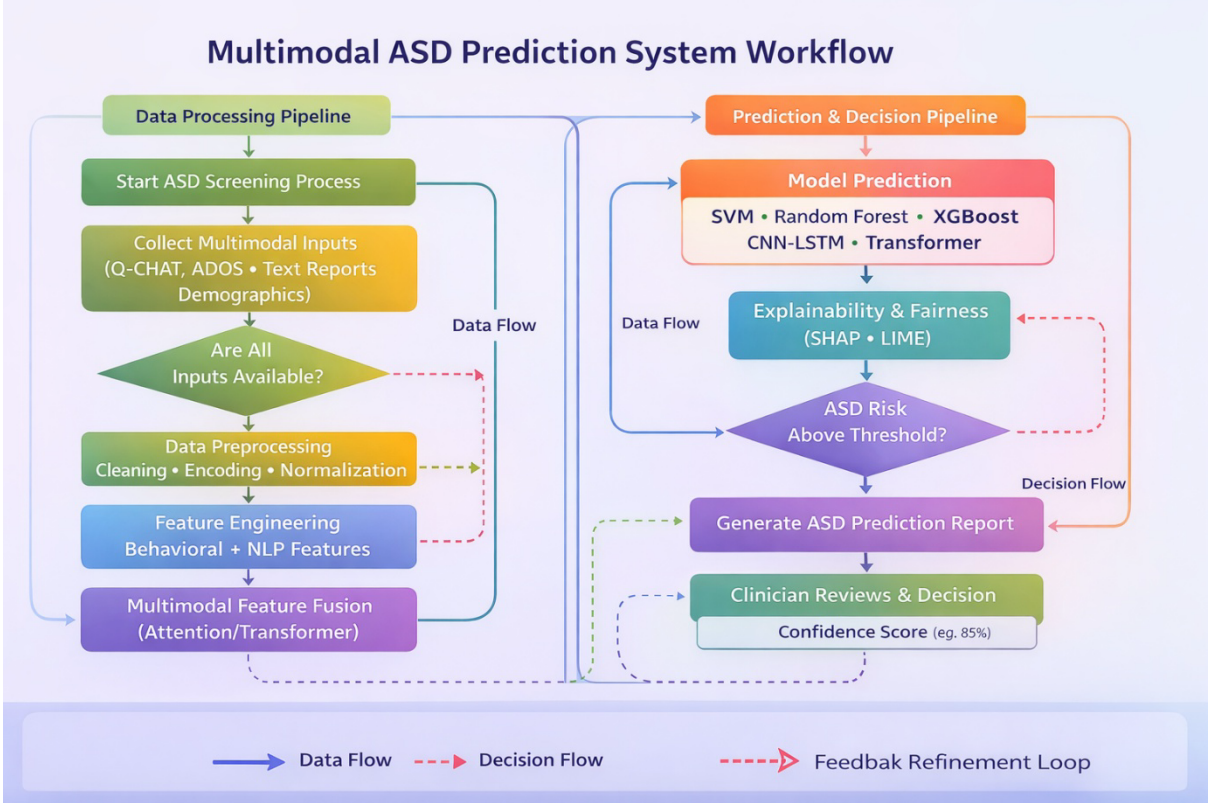

FIGURE 1 PROPOSED MULTIMODAL ASD PREDICTION FRAMEWORK ORGANIZED INTO A TWO-COLUMN WORKFLOW SHOWING DATA ACQUISITION, PREPROCESSING, MULTIMODAL FUSION, PREDICTION, EXPLAINABILITY, AND CLINICAL DECISION-MAKING STAGES.

4.1 Application Use Case: Early ASD Screening Support System

The implemented framework was evaluated within a simulated early ASD screening support scenario representing real clinical workflows. In this application use case, caregivers or healthcare practitioners initiate screening by providing behavioural questionnaire responses, demographic information, and caregiver observations describing social and communication behaviour. These multimodal inputs are submitted to the proposed system through a digital screening interface.

Following data submission, the preprocessing module standardizes behavioural and textual information before feature engineering extracts behavioural indicators and linguistic patterns associated with ASD traits. The multimodal fusion module integrates structured behavioural variables with textual representations to generate a unified feature space. Machine learning and deep learning models then estimate an ASD risk probability accompanied by explainability outputs generated using SHAP and LIME techniques.

If the predicted ASD risk exceeds a predefined threshold, the system produces an interpretable prediction report highlighting influential behavioural and linguistic features. Clinicians subsequently review the generated report to support diagnostic decision-making and early intervention planning. This application scenario demonstrates how the proposed framework transitions from experimental modelling to practical decision-support deployment, and the performance results presented in Section 5 are obtained under this operational workflow.

Data preprocessing and feature engineering were performed to ensure consistency, reduce noise, and enhance model learning. Structured behavioural and demographic variables were examined for missing values, with numerical features imputed using mean or median values depending on distribution characteristics and categorical variables completed using mode-based or k-nearest neighbour imputation. Categorical attributes such as gender and ethnicity were encoded using one-hot encoding to avoid artificial ordinal relationships. Continuous variables were normalized using z-score normalization, defined as

$$x' = \frac{x - \mu}{\sigma}$$

where $\mu$and σdenote the mean and standard deviation of the feature. For unstructured textual data, naïve natural language processing techniques were applied to preserve interpretability and computational efficiency. Text preprocessing included lowercasing, tokenization, removal of punctuation and stop words, and lemmatization. Feature extraction employed term frequency–inverse document frequency (TF-IDF), expressed as

$$\text{TF-IDF}(t, d) = tf(t, d) \log\left(\frac{N}{df(t)}\right)$$

where $tf(t, d)$is the term frequency, $df(t)$is the document frequency, and $N$is the total number of documents. Additional linguistic features, including lexical diversity and sentiment polarity, were extracted to capture ASD-related communication patterns.

A diverse set of machine learning models was implemented to evaluate the effectiveness of the multimodal framework. Classical models such as Naïve Bayes, Support Vector Machines, Random Forest, and XGBoost were selected for their robustness on tabular and high-dimensional data and compatibility with naïve NLP features. Advanced architectures, including Long Short-Term Memory networks and attention-based multimodal models, were employed to capture temporal dependencies and cross-modal interactions. Model training was conducted using stratified k-fold cross-validation to preserve class distributions and reduce variance, with the average cross-validation score computed as

$$CV = \frac{1}{k}\sum_{i=1}^{k} \text{Score}_i$$

To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) and focal loss were applied, with focal loss defined as

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t))$$

All experiments were conducted in a controlled Python-based environment using established machine learning, deep learning, and explainable AI libraries, ensuring reproducibility, fairness evaluation, and practical deploy ability.

5. Performance Evaluation and Results

The performance of the proposed multimodal expert system for autism spectrum disorder (ASD) prediction was assessed using a comprehensive set of classification and reliability metrics to ensure accuracy, robustness, and clinical suitability. Given the sensitive nature of ASD screening and the presence of class imbalance in real-world datasets, evaluation focused not only on overall accuracy but also on class-wise and threshold-independent performance. Standard metrics included accuracy, precision, recall, and F1-score. Accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FNdenote true positives, true negatives, false positives, and false negatives. Precision and recall, which are particularly important for minimizing missed ASD cases, are computed as

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN},$$

and the F1-score is given by

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

To provide threshold-independent assessment, ROC-AUC was reported, while PR-AUC and balanced accuracy were included to account for class imbalance. Model validation was conducted using stratified k-fold cross-validation, with performance scores averaged across folds to ensure stability and reproducibility.

A comparative performance analysis was conducted across classical and advanced machine learning models using early-fusion multimodal features derived from behavioural screening data, caregiver textual reports, and demographic metadata. Classical models such as Naïve Bayes, Support Vector Machines (SVM), Random Forest,

and XGBoost demonstrated competitive performance. Naïve Bayes performed particularly well when combined with TF-IDF textual features, highlighting the effectiveness of naïve NLP representations in capturing ASD-related linguistic patterns. SVM and Random Forest models effectively learned non-linear relationships in the multimodal feature space, while XGBoost achieved consistent performance due to its regularization and ensemble learning capabilities, making it robust to high-dimensional and noisy inputs. Deep learning architectures, including LSTM and attention-based multimodal networks, outperformed classical models on most metrics, especially recall and ROC-AUC. The attention-based fusion model achieved the best overall performance by dynamically weighting informative modalities, although it required higher computational resources and offered reduced interpretability, reflecting a trade-off between predictive performance and practical deploy ability.

The proposed system was evaluated separately on Toddler and Adult ASD screening datasets to examine robustness across age groups and diagnostic contexts. On the Toddler dataset, unimodal models based on behavioural and demographic features achieved moderate accuracy and recall, whereas multimodal integration produced substantial gains, particularly in recall and ROC-AUC, which are critical for early childhood screening. The attention-based fusion model demonstrated strong robustness to noisy or incomplete inputs, maintaining consistent performance across validation folds. On the Adult dataset, performance improvements were also observed with multimodal integration, though to a lesser extent due to greater behavioural heterogeneity and variability in self-reported data. Ensemble and deep learning models exhibited more stable performance than classical approaches. Fairness-aware evaluation across both datasets revealed reduced demographic performance disparities compared to unimodal baselines.

An ablation analysis further highlighted the contribution of individual modalities. Behavioural features alone provided a strong baseline, reflecting the clinical validity of standardized screening tools, but achieved lower recall and ROC-AUC than the full multimodal system. The inclusion of caregiver textual data significantly improved recall and PR-AUC, emphasizing the importance of narrative context in identifying subtle behavioural cues. Demographic metadata, while not a strong predictor in isolation, contributed to contextual interpretation and fairness-sensitive assessment. Attention-guided fusion consistently outperformed simple feature concatenation by selectively emphasizing informative modalities, confirming that multimodal integration yields the most robust, interpretable, and ethically aligned ASD prediction performance.

6. Fairness and Explainability Analysis

6.1 Fairness Metrics and Bias Assessment

Fairness is a critical requirement for artificial intelligence systems deployed in healthcare, particularly for conditions such as autism spectrum disorder (ASD), where demographic disparities in diagnosis and access to care have been widely reported. Predictive models trained on historical data may unintentionally learn and amplify existing biases related to gender, ethnicity, or socioeconomic status. Therefore, the proposed multimodal expert system incorporates a systematic fairness evaluation framework to assess and monitor demographic bias.

Fairness assessment is conducted using established group-based metrics across sensitive attributes such as gender and age group. One primary metric employed is **Demographic Parity Difference (DPD)**, which evaluates whether the predicted positive rate is comparable across demographic groups:

$$DPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)$$

where $\hat{Y}$ denotes the predicted ASD label and A represents a sensitive attribute. A DPD value close to zero indicates fair treatment across groups. Additionally, **Equal Opportunity Difference (EOD)** is used to measure disparities in true positive rates:

$$EOD = TPR_{A=0} - TPR_{A=1}$$

This metric is particularly relevant in ASD prediction, as unequal sensitivity across groups could result in delayed identification for certain populations. Fairness evaluation is conducted on both Toddler and Adult datasets, and results are reported alongside standard performance metrics to ensure balanced assessment. The analysis reveals that unimodal models often exhibit higher demographic disparities, whereas multimodal integration reduces bias by capturing a more comprehensive representation of behavioural and contextual factors.

6.2 Bias Mitigation Techniques and Impact

When demographic disparities are identified, bias mitigation strategies are applied to reduce unfair outcomes while preserving predictive performance. The expert system employs both data-level and algorithm-level interventions. At the data level, instance reweighting adjusts the influence of samples from underrepresented or disadvantaged groups during training, promoting balanced decision boundaries. At the algorithm level, fairness-

aware learning is implemented through cost-sensitive training and regularization constraints. For deep learning models, focal loss combined with group-specific weighting emphasizes misclassified instances from minority groups. The effectiveness of these interventions is assessed by comparing fairness metrics before and after mitigation. Results show consistent reductions in demographic parity and equal opportunity gaps with minimal impact on accuracy and ROC-AUC. In some cases, fairness-aware training improves generalization by reducing overfitting to dominant groups, demonstrating that fairness and performance can be jointly optimized.

6.3 Explainable AI Techniques

While fairness ensures equitable outcomes across groups, explainability addresses transparency and trust at the individual prediction level. In clinical decision support, practitioners require insight into why a model produces a particular prediction rather than relying on opaque outputs. To this end, the proposed system integrates multiple explainable AI (XAI) techniques that provide both global and local interpretability.

6.3.1 Global Interpretability using SHAP

Global interpretability is achieved using **SHapley Additive explanations (SHAP)**, which quantify the contribution of each feature to the model's overall predictions. SHAP values are grounded in cooperative game theory and represent the average marginal contribution of a feature across all possible feature subsets. For a model f, the SHAP value for feature $i$is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)]$$

Global SHAP analysis reveals that behavioural screening scores consistently dominate prediction outcomes, followed by textual linguistic features and selected demographic attributes. This aligns with clinical expectations and validates the model's reliance on meaningful diagnostic cues rather than spurious correlations.

6.3.2 Local Interpretability using LIME

To explain individual predictions, **Local Interpretable Model-Agnostic Explanations (LIME)** are employed. LIME approximates the complex model locally using a simpler, interpretable surrogate model around a specific instance. For a given input x, LIME solves:

$$\arg\min_{g \in G} \; L(f, g, \pi_x) + \Omega(g)$$

where $g$is an interpretable model, $\pi_x$defines locality around x, and $\Omega(g)$penalizes model complexity. Local explanations highlight which features most strongly influenced a specific ASD prediction, such as limited eye contact indicators or linguistic patterns in caregiver reports. These explanations provide actionable insights for clinicians and caregivers, supporting informed decision-making.

6.3.3 Counterfactual Explanation Analysis

Counterfactual explanations further enhance interpretability by identifying minimal changes to input features that would alter the model's prediction. Formally, a counterfactual $x'$is defined such that:

$$\hat{Y}(x) \neq \hat{Y}(x') \text{and} \parallel x - x' \parallel \text{ is minimal}$$

In the context of ASD prediction, counterfactuals illustrate how changes in behavioural scores or linguistic indicators could shift an individual from high-risk to low-risk classification. These explanations are particularly valuable for early intervention planning, as they highlight modifiable factors rather than static attributes.

Overall, the integration of fairness auditing and explainable AI transforms the proposed system from a black-box predictor into a transparent, ethically aligned decision-support tool. By jointly addressing equity and interpretability, the system enhances clinical trust and supports responsible deployment in real-world ASD screening environments.

7. Discussion

The experimental results indicate that the proposed multimodal expert system achieves robust and consistent performance across both Toddler and Adult ASD screening datasets. By integrating behavioural screening data, caregiver textual reports, and demographic information, the system significantly improves accuracy, recall, and ROC-AUC compared to unimodal approaches. The improvement in recall is particularly clinically relevant, as reducing false negatives is essential for early ASD identification and timely intervention. Multimodal learning enables a richer and more comprehensive representation of ASD-related characteristics by combining quantitative behavioural indicators with qualitative contextual insights from caregiver narratives. The attention-based fusion mechanism further enhances reliability by prioritizing the most informative modalities under noisy or incomplete

data conditions, while interpretable outputs and confidence scores strengthen the system's role as a clinically meaningful decision-support tool.

Compared with existing ASD prediction studies, the proposed approach offers clear methodological and ethical advantages. Many prior works rely on unimodal data sources that perform well in controlled settings but lack robustness in real-world applications. In contrast, this study integrates complementary modalities while explicitly addressing fairness and interpretability. Unlike recent deep learning models that prioritize accuracy at the expense of transparency, the use of interpretable fusion strategies and naïve NLP techniques achieves a balanced trade-off between performance and explainability. The systematic assessment of demographic bias and application of bias mitigation further distinguish this work. Practically, the system provides a scalable and deployable solution for early ASD screening, suitable for primary care and community settings. Its explainable, fairness-aware design supports clinical trust, informed decision-making, and equitable, cost-effective early intervention.

8. Limitations and Future Work

Despite the strong performance of the proposed multimodal expert system, several limitations and future research directions should be acknowledged. The evaluation relies mainly on publicly available ASD screening datasets, which may not fully represent the diversity and complexity of real-world clinical populations due to demographic skewness, limited geographic coverage, and class imbalance, potentially affecting generalizability. Although the framework integrates behavioural, textual, and demographic modalities, the availability and quality of caregiver reports vary across healthcare settings, and complete multimodal data may not always be accessible in practice. Performance may therefore degrade when key inputs or longitudinal information are missing. Additionally, while naïve natural language processing techniques enhance interpretability and efficiency, they may not fully capture deeper semantic patterns in complex clinical language. Future work should focus on validating the system using larger, more diverse, and longitudinal datasets, integrating additional modalities such as speech or sensor data, and exploring hybrid NLP and missing-modality-aware architectures. Large-scale clinical trials are essential to assess real-world usability, fairness, and clinical impact.

9. Conclusion

This study developed and validated a systematically designed multimodal expert system for autism spectrum disorder (ASD) prediction, addressing key limitations of traditional diagnostic approaches and unimodal AI models. By integrating behavioural screening data, caregiver textual reports, and demographic metadata, the framework provided a comprehensive and context-aware representation of ASD-related characteristics. Experimental results on Toddler and Adult ASD datasets demonstrated that multimodal integration consistently outperformed unimodal baselines in terms of accuracy, recall, and ROC-AUC, highlighting the value of combining quantitative behavioural indicators with qualitative linguistic cues for early risk prediction. The adoption of an attention-based fusion strategy improved robustness under noisy or incomplete inputs. Furthermore, fairness analysis revealed reduced demographic bias, while bias mitigation techniques preserved predictive performance. The integration of explainable AI methods enhanced transparency, interpretability, and clinical trust, demonstrating the feasibility of an ethical, scalable, and clinically acceptable decision-support system for early ASD screening.

References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
- Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., … Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years — Autism and Developmental Disabilities Monitoring Network. *MMWR Surveillance Summaries*, 67(6), 1–23. https://doi.org/10.15585/mmwr.ss6706a1
- Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *British Journal of Psychiatry*, 161(6), 839–843. https://doi.org/10.1192/bjp.161.6.839

- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females. *Journal of Autism and Developmental Disorders*, 31(1), 5–17. https://doi.org/10.1023/A:1005653411471
- Bertini, E., Cirillo, D., Oğuz, B., Santucci, G., & Silva, C. (2019). Interpretable machine learning models for decision support. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300831
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatric Research*, 65(6), 591–598. https://doi.org/10.1203/PDR.0b013e31819e7203
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
- Lime: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule* (2nd ed.). Western Psychological Services.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., … Dietz, P. M. (2020). Prevalence of autism spectrum disorder among children aged 8 years. *MMWR Surveillance Summaries*, 69(4), 1–12. https://doi.org/10.15585/mmwr.ss6904a1
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rosenblatt, M. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. https://doi.org/10.1037/h0042519
- Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426–1448. https://doi.org/10.1017/S0033291719000151
- Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *ACM Conference on Equity and Access in Algorithms*, 1–9. https://doi.org/10.1145/3465416.3483305
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.
- Zhang, Z., & Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31, 8778–8788.