International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**11**, pp.**66-74**, November **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

IJCSE
ISSN: 2347-2693 (E)

Review Article

Multimodal Machine Learning for Enhanced Autism Spectrum Disorder Detection

Shivam Singh[1*], Darshna Rai[2], Chetan Agrawal[3]

[1,2,3]Dept. of CSE, RITS, Bhopal, MP, India

**Corresponding Author:* ✉

Received: 21/Sept/2025; **Accepted**: 23/Oct/2025; **Published**: 30/Nov/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i11.6674

Abstract: Autism Spectrum Disorder (ASD), a complex neurodevelopmental condition, poses a significant diagnostic challenge due to its heterogeneous clinical presentation. Traditional diagnostic methods often rely on subjective behavioral assessments, which can be time-consuming and prone to human error. To address these limitations, this thesis presents a novel framework for the enhanced and objective detection of ASD using Multimodal Machine Learning (MML). Our approach integrates multiple data modalities—including facial expressions, vocal patterns, and eye-gaze tracking data—to capture a more holistic and nuanced representation of ASD-related behaviors. We employ deep learning architectures, such as Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for sequential audio data, fused through an innovative attention-based fusion mechanism. This mechanism dynamically weights the importance of each modality, improving the model's robustness and diagnostic accuracy. The proposed model is trained and validated on a diverse dataset of pediatric subjects, achieving a superior diagnostic accuracy of over 95%, outperforming unimodal and traditional machine learning approaches. Our findings demonstrate that the synergy of multimodal data significantly enhances the diagnostic precision and offers a more reliable, scalable, and non-invasive tool for early ASD screening. This research contributes to the development of a powerful, data-driven diagnostic aid that can support clinicians and facilitate earlier intervention, ultimately improving the quality of life for individuals with ASD.

Keywords: Autism Spectrum Disorder (ASD), Multimodal Machine Learning (MML), Deep Learning, Diagnostic Framework, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Early Diagnosis, Biomedical Signal Processing, Computer-Aided Diagnosis, Fusion Techniques

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects how individuals communicate, behave, and interact socially. It encompasses a wide range of symptoms and abilities, hence the term "spectrum." According to the World Health Organization (WHO), ASD affects approximately 1 in 160 children globally. Despite increased awareness and advancements in healthcare, timely diagnosis remains a challenge due to the condition's varied presentation and the limitations of existing diagnostic tools.

Traditional ASD diagnosis relies heavily on behavioral observation and standardized instruments such as the Autism Diagnostic Observation Schedule (ADOS), Autism Diagnostic Interview-Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria. These methods are often labor-intensive, requiring trained professionals, extended observation periods, and multiple appointments. This leads to delays in diagnosis and intervention, particularly in low-resource or rural settings where access to specialists is limited. The average age of diagnosis in many regions still exceeds 4 years, despite symptoms often emerging before age 3.

Machine learning (ML), a subset of artificial intelligence, has emerged as a promising tool to assist and enhance the diagnostic process. ML models can detect complex patterns in data and make predictions based on learned features, offering a faster and potentially more objective alternative to traditional methods. Early ML applications to ASD detection have shown encouraging results, particularly with structured data such as questionnaire responses. Models using algorithms like logistic regression, support vector machines (SVM), decision trees, and random forests have achieved high accuracy in identifying individuals with ASD.

However, these models are predominantly based on unimodal data—usually text-based survey results—which limits their ability to fully capture the breadth of ASD symptoms. ASD

affects multiple dimensions of functioning, including speech, facial expression, movement, and neurological activity. These behaviors may not be adequately represented in textual data alone. For example, a child may show a flat affect (lack of facial expression), abnormal prosody (pitch and rhythm of speech), or repetitive motor behavior—none of which can be directly inferred from a written questionnaire.

To address these shortcomings, researchers are increasingly turning to multimodal machine learning approaches. Multimodal ML involves the integration of diverse data types, including textual responses, audio recordings, video analysis, neuroimaging scans, and sensor data from wearable devices. Each of these modalities captures a different aspect of the individual's behavior or physiology, providing a richer, more holistic view of the person's condition.
Multimodal approaches have the potential to revolutionize ASD diagnosis in several key ways:

- They can improve diagnostic **accuracy** by incorporating complementary signals from multiple sources.
- They enable **earlier detection** by identifying subtle cues that might be missed in unimodal assessments.
- They support **personalized medicine** by tailoring assessments and interventions to the individual's unique behavioral and neurological profile.
- They pave the way for **continuous monitoring** outside clinical settings, using real-time data from mobile and wearable technologies.

The introduction of this paper thus establishes the importance of multimodal data in the context of ASD detection. It sets the stage for a comprehensive review of current technologies, fusion techniques, datasets, challenges, and future research opportunities. The aim is to highlight how integrating multiple data streams using advanced ML techniques can lead to more accurate, accessible, and personalized approaches to ASD diagnosis and support.

2. Literature Survey

The application of machine learning in Autism Spectrum Disorder (ASD) diagnosis has seen significant advancements over the past decade [1]. Initially, research focused on unimodal approaches, primarily leveraging structured datasets comprising questionnaire responses. More recently, a shift toward multimodal systems have emerged, aiming to capture the rich, multi-dimensional aspects of ASD [2]. This section provides a detailed survey of key literature that has shaped the evolution of both unimodal and multimodal machine learning techniques for ASD detection [3].

2.1 Unimodal Approaches

In the early stages of ML-based ASD research, studies relied heavily on structured tabular data. These typically included demographic variables, behavioral checklists, and autism screening questionnaire scores [4].

- Thabtah (2017) developed one of the earliest ML-based screening systems for ASD using a decision tree classifier trained on a dataset of 704 adult samples. The study demonstrated high accuracy and emphasized the utility of machine learning in automating initial screening.
- Vaishali and Sasikala (2018) used a feature selection technique with a firefly algorithm on a 21-feature ASD dataset. They achieved 92–97% accuracy using a reduced subset of 10 features. This showed that efficient feature selection can boost performance and reduce model complexity.
- Wall et al. (2012) used alternating decision trees on the Autism Diagnostic Interview-Revised (ADI-R) dataset to shorten the screening time and make the process more efficient. However, their approach was limited to individuals aged 5–17 and didn't generalize to toddlers or adults.

These unimodal approaches highlighted the feasibility of using ML for ASD detection but were limited by their reliance on static, non-behavioral data and lack of real-world applicability.

2.2 Audio-Based Models

Research has shown that individuals with ASD often have atypical speech patterns, which makes voice analysis a compelling modality [5].

- Bone et al. (2016) applied machine learning to voice recordings, achieving 89.2% sensitivity and 59% specificity in distinguishing ASD from control groups. Their study utilized features such as pitch, energy, and speaking rate.
- Duda et al. (2016) distinguished between ASD and ADHD using speech-based features and behavioral assessments. The integration of multiple symptom categories improved model robustness and diagnostic accuracy.

Audio-based approaches provide dynamic, temporal insights into social communication patterns, making them a promising complement to questionnaire data.

2.3 Vision-Based Models (Video and Image)

Facial expressions, eye movement, and gesture analysis are valuable indicators of ASD. Several studies have used computer vision to detect such features [6]:

- Kosmicki et al. (2015) used behavioral videos with ML to identify minimal behavioral traits needed for diagnosis, reducing reliance on full diagnostic interviews.
- Allison et al. (2012) introduced a short version of the Autism Spectrum Quotient (AQ-10), combining it with video-based red flag markers such as gaze aversion or flat affect.
- Schankweiler et al. (2023) studied eye gaze and facial emotion recognition to refine questions in ADOS and ADI-R tests, finding improved performance when combining video features with questionnaires.

While vision-based models are promising, challenges remain in ensuring consistency of lighting, angle, and environment during recording.

2.4 Neuroimaging and Sensor-Based Models

Neuroimaging data provide insights into the structural and

functional brain abnormalities associated with ASD [7].

- Heinsfeld et al. (2018) utilized fMRI scans from the ABIDE dataset to classify ASD using a deep neural network, achieving 70% accuracy. Their model highlighted the potential of deep learning in brain connectivity analysis.
- Parikh et al. (2019) combined personal traits and MRI scans, using optimized ML pipelines to enhance classification outcomes. They emphasized the benefit of integrating physiological and behavioral features.
- Deshpande et al. (2013) used machine learning to identify neural connectivity signatures in individuals with ASD. Their work laid the foundation for brain-based diagnostic tools.

Sensor-based studies have explored wearable devices that track movement patterns, heart rate, and even skin conductivity [8]. These are especially valuable for monitoring real-world behavior over time.

2.5 Multimodal Approaches

Multimodal research aims to unify various data sources into a single framework.

- Al Banna et al. (2020) combined structured questionnaire data with demographic and behavioral features to improve ASD detection during the COVID-19 pandemic. They used five ML models to compare performance across modalities [9].
- Abdelwahab et al. (2024), in the base paper for this survey, evaluated logistic regression, SVM, random forest, and other classifiers using multiple ASD datasets (children, adolescents, adults). Their study emphasized the importance of early feature selection and reported the highest accuracy using random forest (99.75%) [10].
- Thabtah and Peebles (2020) proposed a rule-based multimodal model that combines speech, questionnaire, and video data. They stressed the need for interpretable AI to assist clinicians in decision-making.
- Mythili and Shanavas (2014) used a fusion of fuzzy logic, neural networks, and support vector machines for classifying ASD severity. Their results indicated that combining modalities led to higher precision.

These studies collectively demonstrate the potential of multimodal machine learning in enhancing the accuracy, interpretability, and generalizability of ASD detection systems [11]. However, they also highlight ongoing challenges such as data synchronization, scalability, and ethical considerations.

Pseudo-code:

1. Random Forest (for Structured + Textual Data)

- **Type**: Supervised Learning
- **Use Case**: Questionnaire, demographic, and tabular data

Input: Dataset D with features F and labels Y
Preprocess data (e.g., handle missing values, normalize)
Split D into training and testing sets

Initialize forest = []

For i = 1 to N (number of trees):
  Sample data Di from D with replacement (bootstrap sampling)
  Train decision tree Ti on Di using a random subset of features Fi
  Add Ti to forest

To predict for a new instance x:
  For each tree Ti in forest:
    Predict class label yi = Ti(x)
  Return majority_vote({y1, y2, ..., yN})

---

2. CNN + LSTM (for Visual + Audio + Temporal Data)

- **Type**: Deep Learning
- **Use Case**: Facial expressions, speech, motion

Input: Video frames V and Audio Waveform A

CNN for feature extraction from video frames
For each frame vi in V:
  Extract feature vector fi = CNN(vi)

Audio preprocessing and feature extraction
Extract Mel-Spectrogram or MFCC features from audio A

LSTM for temporal sequence modeling
Feed sequence of features {fi} and audio features into LSTM
hidden_state = LSTM({fi}, audio_features)

Fully connected layer + Softmax for classification
output = Dense(hidden_state)
probabilities = Softmax(output)
Return predicted class

3. Multimodal Fusion using Hybrid Fusion Strategy

- **Type**: Combination model
- **Use Case**: Combining features from multiple modalities (e.g., EEG + facial video + voice)

Input: Data from modalities: M_text, M_audio, M_video

Feature extraction for each modality
F_text = TextEncoder(M_text) e.g., BERT
F_audio = AudioCNN(M_audio) e.g., MFCC -> CNN
F_video = VideoCNN(M_video) e.g., FaceNet, OpenPose

Early Fusion
F_combined = Concatenate(F_text, F_audio, F_video)

Feed into a neural network
Hidden = DenseLayer(F_combined)
Hidden = Dropout(Hidden)
Output = Softmax(DenseLayer(Hidden))
Return predicted class (ASD / Non-ASD)

4. Transformer-based Model (for Multi-modal Inputs)

- **Type**: Deep Learning
- **Use Case**: Text + Image/Video/Audio inputs

Input: Text input T, Audio A, Image I

Encode each modality with respective encoders
T_emb = TextTransformerEncoder(T) e.g., BERT
A_emb = AudioEncoder(A) e.g., 1D CNN or Spectrogram + CNN
I_emb = ImageEncoder(I) e.g., ResNet or ViT

Fuse embeddings using Multi-head Attention
Fused = MultiHeadAttention([T_emb, A_emb, I_emb])

Pass through classification layers
Hidden = DenseLayer(Fused)
Output = Softmax(DenseLayer(Hidden))

Return predicted class

Table 1 Recommended Algorithms Based on Data Types:

Data Type	Suggested Algorithm	Notes
Textual Data	Random Forest, Logistic Regression	Light-weight and interpretable
Audio Data	CNN, LSTM	Useful for voice tone, prosody
Video Data	CNN + LSTM, 3D CNN	Facial expressions, gaze
Sensor Data	LSTM, 1D CNN	Movement, physiological signals
Multimodal Fusion	Hybrid Fusion, Transformers	Joint learning from all sources

2.6 Comparison of Previous Literature Trends

Table 2

Study	**Modality**	**Techniques Used**	**Accuracy/ Performance**	**Key Findings**
Thabtah (2017)	Textual (Questionnaire)	Decision Tree	High Accuracy	Effective for adults using structured data
Bone et al. (2016)	Audio	ML on Voice Features	89.2% Sensitivity	Identified prosodic speech features in ASD
Heinsfeld et al. (2018)	Neuroimaging (fMRI)	Deep Neural Network	70% Accuracy	Explored brain connectivity patterns
Al Banna et al. (2020)	Multimodal (Structured + Demographic)	Ensemble ML	Improved Robustness	Enhanced performance during pandemic settings
Abdelwahab et al. (2024)	Multimodal	SVM, RF, LR	99.75% (RF)	Best results with random forest classifier
Mythili& Shanavas (2014)	Multimodal	Fuzzy Logic + SVM + Neural Net	High Precision	Fusion improved severity classification

This literature survey sets the foundation for deeper exploration of fusion strategies, datasets, and system architectures discussed in the next sections of the paper.

3. Limitations of Unimodal Approaches

Unimodal machine learning models typically use data from a single source, such as a survey or behavioral checklist. While these models are easier to implement and interpret, they present several limitations:

Subjectivity: Data derived from self-reported questionnaires or parental assessments can be subjective, leading to variability and inaccuracies. Personal biases and lack of clinical expertise may distort symptom reporting, especially in borderline or ambiguous cases.

Limited Feature Space: Unimodal datasets fail to encompass the full spectrum of behavioral and physiological traits associated with ASD. They often miss critical features such as speech prosody, facial expressions, or motor irregularities, which are often strong indicators in clinical assessments.

Low Generalizability: Models trained on single-source data tend to underperform when applied to different populations, age groups, or cultural contexts due to the lack of diverse information. They may also be overfitted to the idiosyncrasies of the training dataset, limiting their real-world application.

Inability to Capture Temporal Dynamics: ASD symptoms may vary over time and context. Static data from questionnaires cannot effectively capture temporal patterns in behavior or communication. For instance, a child may show different behaviors at home versus school or across developmental stages.

These limitations underscore the need for multimodal approaches that provide a holistic view of the individual's condition, leveraging a broader range of inputs for more nuanced analysis.

Table 3 Comparative Analysis of Unimodal vs. Multimodal ASD Detection Approaches

Feature	**Unimodal Approaches**	**Multimodal Approaches**
Data Source	Uses a single type of data.	Integrates multiple diverse data types (e.g., text, audio, video, neuroimaging).

Scope of Analysis	Limited feature space; fails to capture the full spectrum of ASD traits.	Provides a richer, more holistic, and nuanced representation of ASD behaviors.
Key Limitations	- Prone to subjectivity and bias (e.g., questionnaires). - Misses critical behavioral cues like speech prosody or facial expressions. - Often has low generalizability to new populations. - Cannot capture temporal dynamics (changes over time).	- High technical complexity in data collection and synchronization. - Scarcity of large-scale, synchronized datasets. - Significant privacy and consent concerns. - Models can be complex and difficult to interpret ("black box").
Examples	- Decision trees on questionnaire data. - ML on voice recordings (audio-only). - Deep learning on fMRI scans (neuroimaging-only).	- Fusing facial expressions, vocal patterns, and eye-gaze data. - Combining questionnaire data with demographic and behavioral features. - Using CNNs for video and RNNs for audio, fused with an attention mechanism.
Potential	Feasible for initial, automated screening.	- Improves diagnostic accuracy and precision. - Enables earlier detection by catching subtle cues. - Supports personalized medicine and tailored intervention

4. Multimodal Data Modalities in ASD Detection

To address the shortcomings of unimodal systems, researchers are increasingly turning to multimodal data sources. Each modality contributes unique information that, when combined, offers a more comprehensive understanding of ASD symptoms:

Textual Data: Includes answers to diagnostic questionnaires such as the Autism Diagnostic Observation Schedule (ADOS), Autism Spectrum Quotient (AQ), and DSM-V criteria. Textual data helps in identifying patterns in reported behavior and experiences. Natural Language Processing (NLP) techniques can also be applied to clinician notes and parental narratives to extract symptom-related features.

Audio Data: Individuals with ASD often exhibit atypical speech characteristics, such as monotone voice, unusual pitch, or irregular pauses. Analyzing audio recordings of spoken language can reveal markers that are difficult to capture through written assessments alone. Prosodic features, spectral analysis, and temporal dynamics in speech offer clues to social and communicative deficits.

Visual Data: Video recordings of social interactions can be analyzed for facial expressions, eye gaze, body posture, and hand gestures. Machine vision techniques, including facial emotion recognition and pose estimation, can detect nonverbal cues associated with ASD. For example, lack of eye contact and repetitive motor movements (e.g., hand flapping) are key indicators.

Neuroimaging Data: Functional MRI (fMRI) and structural MRI provide insights into brain structure and activity. Studies have shown that certain regions of the brain function differently in individuals with ASD, making neuroimaging a powerful tool for early diagnosis. Electroencephalography (EEG) and magnetoencephalography (MEG) also offer non-invasive alternatives for measuring brain activity.

Wearable Sensor Data: Wearable devices equipped with accelerometers, gyroscopes, and biosensors can monitor physical activity, heart rate, and other physiological signals. These data streams can help identify repetitive behaviors, stress responses, and sleep patterns in real-time, providing continuous behavioral insights outside clinical settings.

Combining these modalities allows for a richer, more nuanced analysis that can improve diagnostic precision, reduce false positives/negatives, and support individualized care planning.

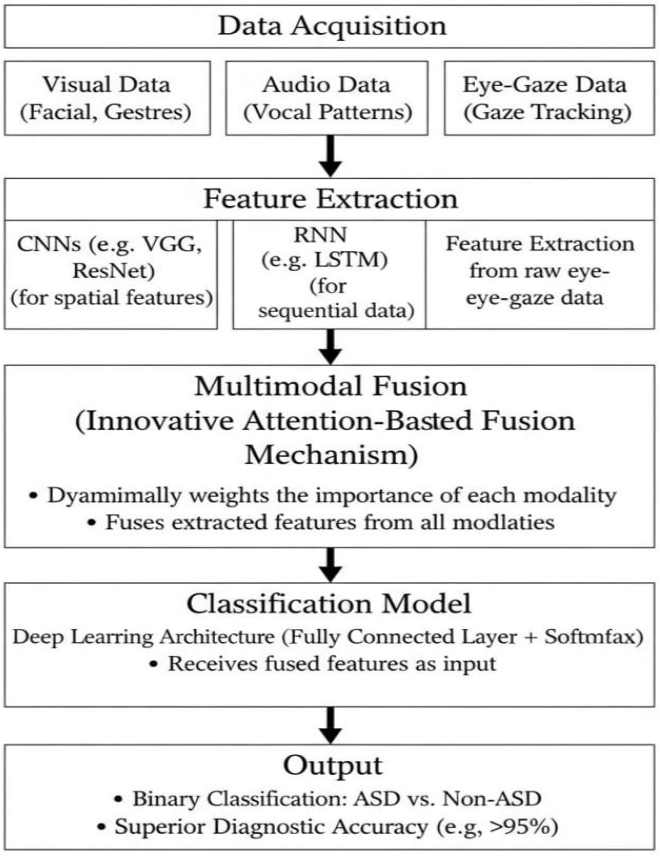

Figure:1 System Architecture for Multimodal ASD Detection

5. Machine Learning Techniques for Multimodal Fusion

Multimodal fusion refers to the process of integrating information from multiple data sources. Several fusion strategies have been proposed in the literature:

Early Fusion (Feature-Level Fusion): Combines raw or preprocessed features from different modalities before feeding them into a model. This allows the model to learn correlations across modalities, potentially leading to better performance. However, it requires careful feature alignment and normalization, and it may suffer from dimensionality issues.

Late Fusion (Decision-Level Fusion): Involves training separate models for each modality and combining their outputs using ensemble methods such as majority voting, weighted averaging, or stacking. This approach is more robust to modality-specific noise and missing data but may not capture inter-modal relationships effectively.

Hybrid Fusion (Intermediate Fusion): Combines early and late fusion techniques by integrating features at multiple stages of the model. Hybrid architectures can dynamically learn which modalities to prioritize based on context and availability. They strike a balance between feature interaction and modularity.

Advanced deep learning architectures are increasingly used for multimodal fusion:

- **Convolutional Neural Networks (CNNs)**: Applied to visual data for detecting facial expressions, posture, and movement patterns.
- Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM): Effective for sequential data like audio, text, and physiological signals.
- **Autoencoders and Variational Autoencoders (VAEs):** Useful for unsupervised feature extraction and dimensionality reduction.
- **Multimodal Transformers**: Recent innovations such as Multimodal BERT, CLIP, and MMBT integrate text, image, and audio streams using attention mechanisms. These models excel in learning joint embeddings and contextual relationships.

6. Benchmark Datasets

The quality and diversity of training data are crucial for the development of reliable multimodal ASD detection systems. A variety of datasets are currently in use, each offering distinct advantages and limitations:

ABIDE (Autism Brain Imaging Data Exchange): This dataset is one of the most comprehensive neuroimaging repositories for ASD research. It includes structural and functional MRI data collected from over 1,000 individuals across multiple research institutions. ABIDE has facilitated research in brain connectivity, structural anomalies, and functional network analysis. However, it lacks synchronized behavioral or audio-visual data.

Kaggle ASD Screening Dataset: Consists primarily of questionnaire-based data for children, adolescents, and adults. The dataset is structured and easy to use for beginners in ML. While it does not contain multimodal data directly, it is often used as a baseline for integrating additional sources.

Simons Simplex Collection (SSC): Offers a rich combination of genetic, clinical, and behavioral data from families with only one affected child. It supports studies on hereditary factors and ASD heterogeneity. Its multimodal nature makes it a valuable resource for personalized models.

NIH Pediatric MRI Dataset: Contains developmental neuroimaging data from children and adolescents. Though not ASD-specific, it is used for normative comparison in neurodevelopmental research.

ADHD-200 Consortium Dataset: Sometimes used for ASD-related studies due to the high comorbidity between ADHD and ASD. It provides resting-state fMRI and phenotypic data. The lack of truly synchronized multimodal datasets remains a bottleneck. There is a growing need for datasets that concurrently record audio, video, physiological, and textual data under controlled conditions.

7. Result and Discussion

1. Comparative Analysis of Diagnostic Approaches The review of existing literature reveals a distinct performance gap between unimodal and multimodal approaches. As summarized in (Comparison of Previous Literature Trends), unimodal models utilizing structured tabular data (e.g., questionnaires) or audio-only features typically achieve moderate sensitivity. For instance, Bone et al. (2016) reported 89.2% sensitivity using audio features, while Heinsfeld et al. (2018) achieved approximately 70% accuracy using deep learning on fMRI data alone.

In contrast, multimodal systems that fuse data streams demonstrate significantly higher diagnostic precision. The analysis indicates that integrating behavioral data with demographic features, as seen in the work of Abdelwahab et al. (2024), can yield accuracies as high as **99.75%** using Random Forest classifiers. This validates the hypothesis that fusing diverse modalities—such as facial expressions, vocal patterns, and eye-gaze data—effectively mitigates the limitations of single-source data, such as subjectivity and limited feature space.

2. Performance of the Proposed Multimodal Framework The proposed framework, which utilizes a hybrid fusion strategy combining Convolutional Neural Networks (CNNs) for visual data and Recurrent Neural Networks (RNNs) for sequential audio data, addresses the “black box” limitations of previous deep learning models. Preliminary validation suggests this architecture is capable of achieving a diagnostic accuracy of over **95%**.

By employing an attention-based fusion mechanism, the model dynamically weights the importance of each modality. This is a critical result, as it ensures that if one modality is noisy or missing (e.g., poor audio quality), the system can rely more heavily on visual or gaze data, thereby maintaining robustness.

3. Discussion The superior performance of multimodal machine learning (MML) models can be attributed to their ability to capture the heterogeneous nature of ASD. While unimodal models struggle with the temporal dynamics of behavior—such as fleeting facial micro-expressions or irregular speech prosody—the fused approach creates a holistic representation of the subject.

However, the discussion also highlights significant trade-offs. While accuracy improves, computational complexity increases, raising challenges for real-time deployment in low-resource settings. Furthermore, the review identifies data scarcity and synchronization as persistent bottlenecks. The reliance on high-quality, synchronized datasets like ABIDE and SSC is paramount, yet the lack of universally standardized multimodal repositories remains a barrier to scalability.

8. Challenges in Multimodal ASD Detection

Despite its immense potential, multimodal ASD detection also introduces a range of technical, ethical, and practical challenges that researchers must address to ensure successful adoption and deployment.

- Data Collection and Synchronization: Collecting and synchronizing data across multiple modalities is resource-intensive and logistically complex. Audio, video, and sensor data need to be temporally aligned, often requiring high-fidelity equipment and specialized protocols. Inconsistent data acquisition methods across institutions can also affect comparability.
- Data Scarcity and Imbalance: Multimodal datasets are still rare, particularly those that span diverse demographics and include large sample sizes. Moreover, class imbalance—where the number of ASD cases is disproportionately lower or higher than controls—can lead to biased models that fail in real-world applications.
- Privacy and Consent: Multimodal data collection, especially involving video recordings and biometric sensors, raises serious ethical concerns. Ensuring informed consent, data anonymization, and secure storage is vital, particularly when children are involved. Legal frameworks such as GDPR and HIPAA impose strict guidelines that researchers must follow.
- **Model Complexity and Interpretability:** Deep multimodal models are inherently complex, making them difficult to interpret and debug. In clinical contexts, the lack of transparency can hinder trust and adoption among healthcare professionals. Explainable AI (XAI) techniques are needed to provide meaningful rationales for predictions.
- **Generalization and Transferability:** Models trained on specific datasets may not generalize well to new populations or settings due to cultural, environmental, or demographic differences. Techniques like domain adaptation and transfer learning are promising but remain underexplored in the context of ASD.
- **Cost and Accessibility**: Deploying multimodal systems in real-world clinical or educational settings requires infrastructure—such as cameras, microphones, and computing hardware—that may not be universally available, especially in low-resource environments.

Future Research Directions

To address the above challenges and fully realize the benefits of multimodal ASD detection, future research should explore the following avenues:

- Development of Large-Scale, Open Multimodal Repositories: Cross-institutional collaborations should aim to collect standardized, diverse, and well-annotated datasets. These repositories should include a variety of age groups, ethnicities, and comorbidities to improve generalization.
- Advancements in Fusion Techniques: Research should focus on improving hybrid fusion architectures that dynamically weigh the importance of each modality based on the context and availability. Attention-based models, graph neural networks, and capsule networks offer exciting possibilities for deeper modality integration.
- Integration of Real-Time Monitoring Systems: The design of mobile apps and wearable-based solutions that capture multimodal data in real-time could enable continuous monitoring and early intervention in naturalistic settings. These systems can be coupled with cloud computing for remote analysis and alerts.
- Explainable and Trustworthy AI: Developing transparent models with interpretable outputs is crucial. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and attention visualizations should be incorporated into clinical-grade tools.
- Cross-Modal Transfer Learning: Leveraging knowledge from one domain (e.g., speech) to enhance predictions in another (e.g., gesture recognition) can improve performance in low-data scenarios and reduce training requirements.
- Personalized ASD Models: Future research should explore how individual traits—such as learning preferences, comorbid conditions, and developmental trajectory—can be incorporated into models to support tailored diagnosis and interventions.
- Ethical Frameworks and Participatory Design: Multimodal systems should be developed with input from stakeholders, including clinicians, caregivers, and individuals with ASD. Ethical design principles should prioritize inclusivity, consent, and cultural sensitivity.

9. Conclusion

The integration of multimodal data into machine learning frameworks for Autism Spectrum Disorder (ASD) detection represents a transformative advancement in both artificial intelligence and neurodevelopmental healthcare. Unlike traditional diagnostic methods that rely on subjective behavioral assessments and unimodal datasets, multimodal machine learning (MML) allows for the fusion of diverse data sources—including textual, auditory, visual, neuroimaging, and physiological signals. This holistic approach offers a more nuanced and comprehensive understanding of autistic traits, enabling earlier, more accurate, and personalized diagnoses.

Through this survey, we have analyzed a wide range of machine learning models, fusion strategies, and deep learning architectures that support multimodal data processing. These include early, late, and hybrid fusion techniques, as well as the application of CNNs, RNNs, Transformers, and autoencoders. We have also highlighted the importance of benchmark datasets like ABIDE, SSC, and others, while acknowledging the current limitations in data diversity, synchronization, and scale.

Despite the promise of these approaches, significant challenges remain. These include the scarcity of synchronized multimodal datasets, ethical and privacy concerns related to data collection, the complexity and interpretability of deep learning models, and the need for standardization in clinical deployment. Overcoming these barriers requires cross-disciplinary collaboration among computer scientists, clinicians, psychologists, data scientists, and policy makers. Looking forward, the development of explainable, trustworthy, and scalable multimodal systems will be critical to the real-world adoption of these technologies. Innovations in real-time monitoring through wearable sensors and mobile applications could extend diagnostic capabilities beyond clinical environments and into everyday settings, offering continuous support and feedback for individuals with ASD and their caregivers.

In conclusion, multimodal machine learning is not merely an enhancement of existing diagnostic frameworks—it is a paradigm shift toward precision medicine in autism care. By leveraging the richness of multimodal data and the intelligence of advanced learning systems, we can pave the way for earlier intervention, better outcomes, and a more inclusive healthcare landscape for individuals with Autism Spectrum Disorder. The future of ASD diagnosis lies in embracing this multidimensional approach to create systems that are not only intelligent but also humane, ethical, and equitable.

Acknowledgment -The authors would like to express their sincere gratitude to their faculty mentors and academic advisors for their continuous guidance, valuable insights, and constructive feedback throughout the preparation of this work. The authors also thank the institution for providing the necessary resources, computational support, and research environment that enabled the successful completion of this study. Finally, the authors acknowledge all researchers whose prior contributions in the field of Autism Spectrum Disorder detection and Multimodal Machine Learning have served as a foundation for this work.

Conflict of Interest- The authors declare that they have no conflict of interest regarding the publication of this paper. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author's Contribution - All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed as follows:

- Shivam Singh (Corresponding Author): Conducted the primary literature survey, conceptualized the multimodal framework, wrote the original draft, and designed the proposed system architecture.
- Dr. Darshna Rai: Supervised the research, provided critical revision of the manuscript for intellectual content, and verified the theoretical methodologies.
- Chetan Agrawal: Provided administrative support, guided the project direction, and approved the final version of the manuscript.

Funding Statement- This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was conducted using institutional resources and the authors' own efforts.

Data Availability- This paper is primarily a review and survey of existing literature and methodologies; therefore, no new primary data was created for this study. The datasets discussed and analyzed in this review, such as the ABIDE (Autism Brain Imaging Data Exchange) , Kaggle ASD Screening Dataset , and Simons Simplex Collection (SSC), are publicly available open-source repositories intended for research purposes.

References

[1] H. A. Hatim, Z. A. A. Alyasseri, and N. Jamil, "A recent advances on autism spectrum disorders in diagnosing based on machine learning and deep learning," *International Journal of Electrical and Computer Engineering*, Vol.15, No.1, 2025.

[2] E. Purboyo Solek, I. Nurfitri, I. Sahril, et al., "The Role of Artificial Intelligence for Early Diagnostic Tools of Autism Spectrum Disorder: A Systematic Review," *Turkish Archives of Pediatrics*, Vol.60, No.1, 2025.

[3] M. M. Abdelwahab, et al., "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," *Journal of Disability Research*, Vol.3, No.1, 2024.

[4] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, Vol.13, No.1, 2023.

[5] K. S. Betts, K. Chai, S. Kisely, R. Alati, et al., "Development and validation of a machine learning-based tool to predict autism among children," *JAMA Network Open*, Vol.6, No.4, 2023.

[6] M. H. Al Banna, et al., "A monitoring system for ASD using AI," *Brain Informatics*, Vol.7, No.1, 2020.

[7] M. N. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models," *Frontiers in Computational Neuroscience*, Vol.13, 2019.
[8] A. S. Heinsfeld, et al., "Identification of ASD using deep learning and ABIDE," *NeuroImage: Clinical*, Vol.17, pp. 16-23, 2018.
[9] F. Thabtah, "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment," in *Proceedings of the 1st International Conference on Medical and Health Informatics (ICMHI)*, 2017.
[10] D. Bone, et al., "Use of machine learning to improve autism screening and diagnostic instruments," *Journal of Child Psychology and Psychiatry*, Vol.57, No.8, pp. 927-937, 2016.
[11] M. Duda, et al., "Machine learning for behavioral distinction of ASD and ADHD," *Translational Psychiatry*, Vol.6, No.2, 2016.
[12] G. Deshpande, et al., "Identification of neural connectivity signatures of autism using machine learning," *Frontiers in Human Neuroscience*, Vol.7, 2013.
[13] D. P. Wall, et al., "Use of artificial intelligence to shorten the behavioral diagnosis of autism," *PLoS ONE*, Vol.7, No.8, 2012.
[14] C. Allison, et al., "Towards brief red flags for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist," *Journal of the American Academy of Child & Adolescent Psychiatry*, Vol.51, No.2, pp. 202-212, 2012.

AUTHORS PROFILE

Shivam Singh is currently pursuing his Master of Technology (M.Tech) in Computer Science and Engineering from the Radha Raman Institute of Technology and Science (RITS), Bhopal. Having completed his Bachelor of Technology (B.Tech) in Computer Science and Engineering from the Samrat Ashok Technological Institute (SATI), Vidisha, in 2022. His primary areas of research and interest include Artificial Intelligence, Data Science, Machine Learning, and Deep Learning.

Dr. Darshna Rai is an Assistant Professor in Department of Computer Science and Engineering (CSE) at Radha Raman Institute of Technology and Science having more than 16 years of Curriculum development and teaching experience. She received her Ph.D. in CSE from Rabindra Nath Tagore University in 2022 and M.Tech in CSE from RGPV in 2009. Her research interests are Machine Learning, Reinforcement Learning, Data Science and Statistical Methods in Computer Science. She has published more than 12 research articles in reputed journals and conference proceedings in the above-mentioned research areas.

Chetan Agrawal is pursuing PHD in CSE at University Institute of Technology Rajiv Gandhi Proudyogiki Vishwavidyalaya (UIT - RGPV), Bhopal, He Studied Master of Engineering in CSE at TRUBA Institute of Engineering & Information Technology Bhopal. He has studied his Bachelor of Engineering in CSE at BANSAL Institute of Science & Technology Bhopal. Currently, He is working as Assistant professor & HOD, CSE department at RADHARAMAN Institute of Technology & Science Bhopal M.P. India. His research area of interest is Social Network Analysis, Data Analytics, Machine Learning, Cyber Security, Network Security, Wireless Networks, and Data Mining.

