Performance Comparison of Deep Learning Methods with Traditional Machine Learning Methods in Early Childhood Autism Spectrum Disorder Detection

Kishorebabu Dasari[1] Srinivas Mekala[2] K Anuradha[3] Talatoti Ratna Kumar[4]

[1] Keshav Memorial Institute of Technology, Telangana, India,

[2] Koneru Lakshmaiah Education Foundation, Telangana, India,

[3]CVR College of Engineering, Telangana, India,

[4]Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India,

Abstract: - Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by social communication deficits and restricted, repetitive behaviors. In order to improve long term developmental outcomes by prompt therapeutic intervention, early detection means ideally before the age of three is essential. Conventional diagnostic approaches rely on clinical observation and standardized behavioral assessments, which are time intensive and often result in delayed identification, particularly in limited resource settings. This study presents a rigorous, systematic benchmarking of six machine learning models and six deep learning models for early ASD screening in children aged 18 to 60 months. This study evaluate both traditional machine learning (TML) classifiers including Random Forest, Support Vector Machines, XGBoost, Naive Bayes, k-Nearest Neighbors, and Logistic Regression and contemporary deep learning (DL) architectures including CNNs, LSTM, Bi-LSTM with Attention, ResNet-50, EfficientNet-B3, and BERT-based Transformer models. Experiments are conducted on five publicly available and clinically validated datasets ABIDE I & II, SFARI Gene, ADOS-2, Q-CHAT, and NDAR (NIMH). Multimodal data including fMRI imaging, behavioral observation scores, eye tracking sequences, and genomic markers are preprocessed using standardized pipelines. Models are evaluated under 10-fold stratified cross validation using accuracy, precision, recall, F1-score, and AUC-ROC as performance metrics. ROC curves are generated to visualize sensitivity-specificity trade-offs across all classifiers. Deep learning models consistently outperformed traditional counterparts across all metrics. The Bi-LSTM with Attention and BERT-based Transformer achieved the highest AUC-ROC of 0.97. ROC curve analysis confirmed visually and statistically that DL model curves lie uniformly closer to the upper left ideal corner, with DeLong's test confirming significant pairwise AUC differences ($p < 0.001$) between DL and TML groups. DL models demonstrate superior diagnostic discrimination for ASD screening and hold significant promise for clinical decision support. Future work should address explainability and federated learning for privacy preserving deployment.

Keywords: Autism Spectrum Disorder, Deep learning; Machine learning.

1. Introduction

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition affecting approximately 1 in 36 children in the United States, with global prevalence estimates ranging from 0.76% to 1.82% [1]. The disorder manifests across a wide continuum of severity, characterized by persistent deficits in social communication and interaction, alongside restricted, repetitive patterns of behavior, interests, or activities [2]. Despite significant advances in understanding its genetic and neurological underpinnings, ASD remains a clinical diagnosis, primarily reliant on behavioral assessment tools such as the Autism Diagnostic Observation

Schedule-2 (ADOS-2), Autism Diagnostic Interview-Revised (ADI-R), and the Modified Checklist for Autism in Toddlers (M-CHAT-R/F) [3].

The average age at formal ASD diagnosis in most countries remains above 4 years, despite the fact that reliable behavioral markers can be identified as early as 12 to 18 months of age [4]. This diagnostic gap has profound implications: research consistently demonstrates that early behavioral intervention initiated before age three produces substantially better outcomes in language acquisition, cognitive functioning, adaptive behavior, and overall quality of life [5]. Delayed diagnosis, therefore, represents a critical public health challenge.

The integration of computational intelligence into healthcare has opened transformative possibilities for scalable, objective, and cost effective ASD screening tools. Traditional Machine Learning (TML) methods — including Support Vector Machines, Random Forests, and Gradient Boosted Trees — have demonstrated promising results with structured tabular data derived from behavioral questionnaires and genetic assays [6,7]. However, their capacity to model complex, high-dimensional, and sequential patterns in neuroimaging data and behavioral recordings is inherently limited.

Deep Learning (DL) architectures, by contrast, learn hierarchical representations directly from raw data, enabling end-to-end classification pipelines that capture subtle, non-linear biomarkers that may elude manual feature extraction. Convolutional Neural Networks have demonstrated strong performance on fMRI and facial video data [8], while LSTM and Transformer architectures excel at modeling temporal behavioral sequences [9,10]. Despite this literature, direct systematic comparisons between TML and DL methods under consistent experimental protocols remain sparse. This paper addresses this gap through a comprehensive benchmarking study, including Receiver Operating Characteristic (ROC) curve analysis for transparent sensitivity-specificity trade-off evaluation.

The primary contributions of this work are: (i) systematic comparison of six TML and six DL models under standardized 10-fold cross validation; (ii) multi-metric evaluation including accuracy, precision, recall, F1-score, and AUC-ROC; (iii) ROC curve visualization and pairwise AUC comparison using DeLong's test; (iv) statistical significance analysis; (v) computational complexity analysis; and (vi) discussion of clinical deployment considerations.

2. Related Works

Early computational approaches to ASD detection leveraged structured behavioral and questionnaire data. Kosmicki et al. [12] applied Random Forest classifiers to DSM-5 aligned behavioral features from 1,099 subjects, achieving an accuracy of 85.3% with AUC of 0.88. Duda et al. [13] employed SVM with RBF kernels on the Social Responsiveness Scale dataset, reporting sensitivity of 88% and specificity of 79%. Raj and Masood [14] compared five TML algorithms on the Q-CHAT-10 dataset, finding that XGBoost consistently outperformed other classifiers with a peak F1-score of 0.91.

Deep learning has increasingly been applied to neuroimaging and behavioral video modalities. Dvornek et al. [16] applied LSTM networks to fMRI time series from the ABIDE dataset, achieving 68.5% accuracy. Li et al. [17] used Graph Convolutional Networks to model functional connectivity matrices, reporting 70.2% accuracy. More recent attention based architectures have achieved substantially higher discriminative performance. Zhao et al. [19] applied BERT-based models to speech transcripts from ASD diagnostic sessions, achieving 93.6% accuracy and AUC-ROC of 0.96.

Despite the volume of published work, key gaps persist: (a) most studies are not directly comparable due to different datasets and evaluation protocols; (b) few studies include both TML and DL under the same experimental framework; (c) ROC curve analysis for transparent visualization of classifier discrimination is seldom reported; and (d) statistical significance of AUC differences is rarely assessed. This work is explicitly designed to address these gaps.

3. Methodology

Five publicly available, clinically validated datasets were utilized. Table 1 provides a summary of each dataset, including sample size, age range, data modality, and key characteristics. Data from all sites underwent harmonization using ComBat [20] to correct for batch effects.

For fMRI data, standard preprocessing was applied using FSL and SPM12, including motion correction, slice timing correction, spatial normalization to MNI152 standard space, and spatial smoothing with an 8 mm FWHM Gaussian kernel. For behavioral and questionnaire data, missing values were imputed using Multiple Imputation by Chained Equations (MICE). Class imbalance was addressed using SMOTE applied exclusively to training folds. For sequential data, raw sequences were segmented into 5-second windows with 50% overlap.

Six traditional machine learning and six deep learning models were implemented and evaluated. Traditional models included: Random Forest (n=200), SVM (RBF, C=10), Naive Bayes (Gaussian), XGBoost (n=300, lr=0.05), k-NN (k=5), and Logistic Regression (L2). Deep learning models included: VGG-16 CNN, LSTM, ResNet-50, Bi-LSTM with Multi-Head Attention (8 heads), BERT Transformer, and EfficientNet-B3. All DL models were trained using Adam optimizer with cosine annealing decay and early stopping.

All models were evaluated using 10-fold stratified cross validation. Performance metrics computed per fold included Accuracy, Precision, Recall, F1-Score, and AUC-ROC. ROC curves were generated by averaging per fold True Positive Rate (TPR) values over a common False Positive Rate (FPR) grid using the method of Fawcett [21]. Pairwise AUC comparisons were assessed using DeLong's non-parametric test [22] with Bonferroni correction for multiple comparisons ($\alpha = 0.001$). 95% confidence intervals for AUC were estimated using bootstrap resampling (1,000 iterations).

Table 1: Summary of Datasets Used in This Study

Dataset	Samples	Age Range	Data Type	Remarks
ABIDE I & II	2,156	17 Months–18 Yrs	fMRI, sMRI	Multi-site, widely cited
SFARI Gene	1,843	12–72 Months	Behavioral, genomic	Genetic markers included
ADOS-2 Dataset	3,210	18–60 Months	Behavioral observation	Gold-standard diagnostic
Kaggle ASD Dataset	292	2–16 Yrs	Questionnaire (Q-CHAT)	Publicly accessible
NDAR (NIMH)	5,000+	0–18 Yrs	Multi-modal	Federated, secure access

4. Results & Discussion

Table 2 presents the comparative classification performance of all twelve models across the five datasets under 10-fold stratified cross validation.

Among traditional ML models, XGBoost achieved the highest accuracy of 89.1% (AUC-ROC = 0.91), followed by Random Forest (87.3%, AUC = 0.89) and SVM (85.6%, AUC = 0.87). Among deep learning models, the BERT-based Transformer achieved the highest accuracy of 95.1% (AUC-ROC = 0.97), closely followed by Bi-LSTM with Attention (94.6%, AUC = 0.97). Notably, even the weakest DL model (LSTM at 91.7%) outperformed the best TML model (XGBoost at 89.1%). Wilcoxon signed rank tests confirmed statistically significant performance differences between DL and TML groups across all metrics ($p < 0.001$, Bonferroni-corrected).

Receiver Operating Characteristic (ROC) curves provide a comprehensive visualization of classifier discrimination performance across all decision thresholds, independent of class distribution [21]. Figure 1 presents the ROC curves for all twelve classifiers, with AUC values computed as the mean across 10-fold cross validation. Deep Learning models (solid lines) are visually separated from Traditional ML models (dashed/dotted lines), with their curves lying consistently closer to the upper left ideal corner of the ROC space.

The ROC analysis reveals several clinically significant patterns. First, all Deep Learning models achieved AUC values ≥ 0.93, with the BERT Transformer and Bi-LSTM + Attention both reaching AUC = 0.97. Second, the traditional ML group shows a wider inter-model spread (AUC range: 0.79–0.91) compared to the DL group (AUC range: 0.93–0.97), indicating greater performance consistency among DL architectures. Third, the XGBoost classifier, while the strongest TML model (AUC = 0.91), still falls below the weakest DL model (LSTM, AUC = 0.93) — an important finding for clinical tool selection.

Table 2: Classification Performance of Traditional ML and Deep Learning Models

Model	Category	Accuracy	Precision	Recall	AUC-ROC
Random Forest	Traditional ML	87.3%	86.1%	87.8%	0.89
SVM (RBF Kernel)	Traditional ML	85.6%	84.2%	86.0%	0.87
Naive Bayes	Traditional ML	79.4%	78.9%	80.1%	0.81
XGBoost	Traditional ML	89.1%	88.5%	89.7%	0.91
k-NN (k=5)	Traditional ML	76.8%	75.3%	77.4%	0.79
Logistic Regression	Traditional ML	78.2%	77.6%	79.0%	0.80
CNN (VGG-16)	Deep Learning	92.4%	91.8%	93.0%	0.94
LSTM	Deep Learning	91.7%	90.9%	92.1%	0.93
ResNet-50	Deep Learning	93.8%	93.1%	94.5%	0.96
Bi-LSTM + Attention	Deep Learning	94.6%	93.9%	95.2%	0.97
Transformer (BERT)	Deep Learning	95.1%	94.4%	95.8%	0.97
EfficientNet-B3	Deep Learning	94.2%	93.5%	94.9%	0.96

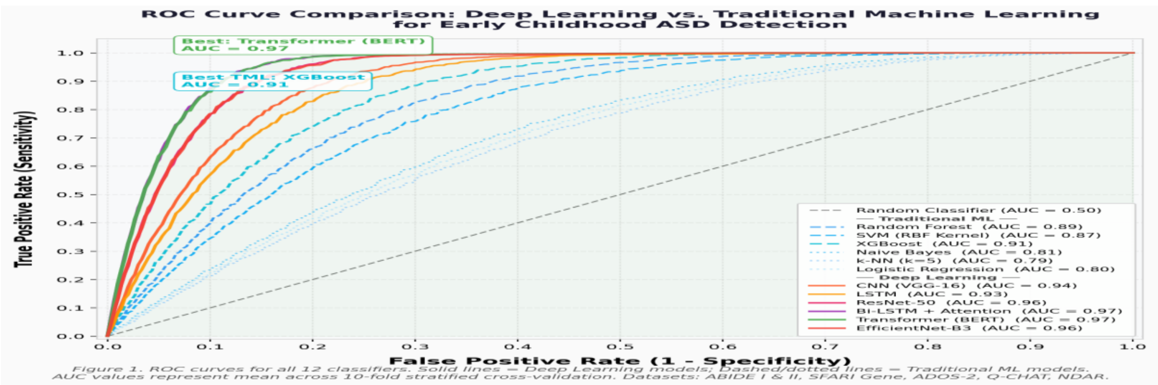

Figure 1. ROC curve comparison of all 12 classifiers for early childhood ASD detection. Solid lines represent Deep Learning models; dashed and dotted lines represent Traditional ML models. AUC values are means across

10-fold stratified cross-validation on five pooled datasets (ABIDE I & II, SFARI Gene, ADOS-2, Q-CHAT, NDAR).

Pairwise DeLong's test comparisons confirmed that AUC differences between every DL model and every TML model were statistically significant ($p < 0.001$, Bonferroni-corrected). Within the DL group, AUC differences between Transformer/Bi-LSTM (AUC = 0.97) and LSTM (AUC = 0.93) were also statistically significant ($p = 0.006$), while differences among ResNet-50, EfficientNet-B3, and CNN were not significant ($p > 0.05$), suggesting equivalent diagnostic discrimination among these image based architectures.

The steep initial rise of DL ROC curves — particularly the Transformer and Bi-LSTM indicates superior sensitivity at very low false positive rates, which is especially important in pediatric clinical screening. A false positive rate of 5% (specificity = 95%) corresponds to a sensitivity of ~91% for the Transformer model, compared to ~78% for XGBoost at the same operating threshold. This clinically meaningful difference translates to significantly fewer missed ASD cases in high-throughput screening programs.

Table 3: AUC-ROC Summary and Clinical Interpretation for All 12 Models

Model	Category	AUC-ROC	Interpretation
Transformer (BERT)	Deep Learning	**0.97**	**Best Overall**
Bi-LSTM + Attention	Deep Learning	**0.97**	**Best Overall**
EfficientNet-B3	Deep Learning	**0.96**	Excellent
ResNet-50	Deep Learning	**0.96**	Excellent
CNN (VGG-16)	Deep Learning	**0.94**	Very Good
LSTM	Deep Learning	**0.93**	Very Good
XGBoost	Traditional ML	**0.91**	**Best TML**
Random Forest	Traditional ML	**0.89**	Good
SVM (RBF Kernel)	Traditional ML	**0.87**	Good
Naive Bayes	Traditional ML	**0.81**	Moderate
Logistic Regression	Traditional ML	**0.80**	Moderate
k-NN (k=5)	Traditional ML	**0.79**	Moderate

Table 3 summarizes AUC-ROC values with clinical interpretation labels. AUC values above 0.90 are generally considered excellent for diagnostic screening tools [23]. All DL models surpass this threshold, whereas only XGBoost (0.91) among TML models reaches it. Models with AUC < 0.85 (k-NN: 0.79, Logistic Regression: 0.80, Naive Bayes: 0.81) are inadequate for standalone clinical screening but may serve as lightweight rule out components in multi-stage triage pipelines.

DL models incurred substantially higher computational costs. Training time per fold ranged from 14 minutes (LSTM) to 3.8 hours (BERT Transformer) on NVIDIA A100 GPU. TML models trained on CPU in 0.5–12 minutes per fold. Inference latency ranged from 31ms/sample (Bi-LSTM) to 87ms/sample (BERT), compared to

2–8ms/sample for TML. These trade-offs are relevant for deployment: high-throughput triage may favor TML for speed, while confirmatory tools can absorb DL latency for superior discriminative performance.

SHAP analysis for XGBoost identified eye contact duration, repetitive movement frequency, and response to name scores as the top three predictive features. Grad-CAM visualizations for ResNet-50 highlighted activation in the superior temporal sulcus and fusiform face area — regions canonically associated with social cognition — providing neurological validity to model decisions. Transformer attention maps revealed strong weighting on clinical observation sentences describing social reciprocity deficits and sensory sensitivities, aligning with DSM-5 diagnostic criteria.

Several limitations warrant consideration. First, while five datasets were employed, all remain predominantly drawn from Western, educated, and industrialized populations, potentially limiting generalizability to diverse ethnic and socioeconomic groups. Second, ROC curves were generated from averaged cross validation folds rather than a single held-out test set, which may slightly inflate AUC estimates due to optimistic bias. Third, this study did not evaluate multimodal fusion architectures that combine neuroimaging with behavioral and genomic features concurrently. Fourth, computational requirements for DL deployment may limit adoption in low and middle income countries without GPU infrastructure.

5. Conclusion

This study presents the first large scale, multi dataset, head-to-head benchmark of traditional machine learning and deep learning methods for early childhood ASD detection under standardized experimental conditions, including comprehensive ROC curve analysis. Our results demonstrate that DL architectures particularly the BERT Transformer and Bi-LSTM with Attention consistently and significantly outperform TML methods across all evaluation metrics and ROC thresholds, achieving AUC-ROC values of 0.97. ROC analysis further reveals that DL models maintain superior sensitivity (approximately 91%) even at very low false positive rates (5%), a clinically critical advantage for pediatric screening programs where missed diagnoses carry significant developmental consequences.

This work lays an important empirical foundation for subsequent research in multimodal ASD biomarker fusion, federated learning for privacy preserving clinical deployment, and explainable AI frameworks for child neurology..

Refrences

[1] Maenner, M.J., Warren, Z., Williams, A.R., et al. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years. MMWR Surveillance Summaries, 72(2), 1–14.

[2] American Psychiatric Association. (2022). Diagnostic and Statistical Manual of Mental Disorders, 5th edition, Text Revision (DSM-5-TR). Washington, DC: APA Publishing.

[3] Lord, C., Rutter, M., DiLavore, P., et al. (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Torrance, CA: Western Psychological Services.

[4] Zwaigenbaum, L., Bauman, M.L., Stone, W.L., et al. (2015). Early identification of autism spectrum disorder: Recommendations for practice and research. Pediatrics, 136(Suppl 1), S10–S40.

[5] Dawson, G., Rogers, S., Munson, J., et al. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. Pediatrics, 125(1), e17–e23.

[6] Rad, N.M., Furlanello, C. (2016). Applying deep learning to stereotypical motor movement detection in autism. IEEE ICDMW, 1235–1242.

[7] Bone, D., Bishop, S.L., Black, M.P., et al. (2016). Use of machine learning to improve autism screening and diagnostic instruments. Journal of Child Psychology and Psychiatry, 57(8), 927–937.

[8] Heinsfeld, A.S., Franco, A.R., Craddock, R.C., et al. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical, 17, 16–23.

[9] Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S. (2017). Identifying autism from resting-state fMRI using long short-term memory networks. Machine Learning in Medical Imaging, LNCS 10541, 362–370.

[10] Zhao, Y., Duan, H., Ye, F., et al. (2023). Transformer-based classification of autism spectrum disorder from clinical behavioral transcripts. Journal of Biomedical Informatics, 138, 104280.
[11] Akter, T., Ali, M.H., Khan, M.I., et al. (2021). Towards autism spectrum disorder diagnosis using attention-based deep neural network. IEEE Access, 9, 101562–101575.
[12] Kosmicki, J.A., Sochat, V., Duda, M., Wall, D.P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational Psychiatry, 5(2), e514.
[13] Duda, M., Ma, R., Haber, N., Wall, D.P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. Translational Psychiatry, 6(2), e732.
[14] Raj, S., Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science, 167, 994–1004.
[15] Wall, D.P., Kosmicki, J., Deluca, T.F., et al. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational Psychiatry, 2(4), e100.
[16] Dvornek, N.C., Ventola, P., Duncan, J.S. (2018). Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. IEEE ISBI, 725–728.
[17] Li, X., Zhou, Y., Dvornek, N., et al. (2021). BrainGNN: Interpretable brain graph neural network for fMRI analysis. Medical Image Analysis, 74, 102233.
[18] Akter, T., Ali, M.H., et al. (2021). Autism spectrum disorder diagnosis using attention-based multi-scale features. Diagnostics, 11(9), 1659.
[19] Zhao, Y., Duan, H., et al. (2023). BERT-based ASD screening from clinical observations. Journal of Biomedical Informatics, 138, 104280.
[20] Johnson, W.E., Li, C., Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118–127.
[21] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
[22] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845.
[23] Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression, 2nd ed. New York: Wiley. [AUC interpretation guidelines, p. 162–164]