Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 28 April 2026
Revised 30 May 2026
Accepted 16 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5): 230–251

KEYWORDS:
Autism Spectrum Disorder, Deep Learning, SHAP, Explainable AI,Early Detection, Stage Classification, CNN-LSTM, Healthcare Analytics

Deep Learning-based Early Autism Spectrum Disorder Detection and Stage Classification Using SHAP Explainable AI Framework

M. Ramesh[1], C. Vijayakumaran[2]

[1]Research Scholar, School of Computing, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
[2]Associate Professor, School of Computing, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
Email: [1]rm3515@srmist.edu.in, [2]vijayakc@srmist.edu.in

ABSTRACT: The Early Autism Spectrum Disorder (ASD) detection based on Deep Learning is critical to allow timely clinical intervention and personalized therapy planning. The proposed study is an explainable deep learning framework to detect early ASD and classify the stage using SHAP (SHapley Additive exPlanations) to improve model transparency and clinical trust. The main aim is to come up with a strong and interpretable model that can effectively determine the presence of ASD and the stages of its severity using heterogeneous data. The methodology is a hybrid deep learning architecture that combines Convolutional Neural Networks (CNN) to extract features and Long Short-Term Memory (LSTM) networks to analyze temporal patterns. The model is trained using real-world behavioral and clinical data, and includes preprocessing, feature normalization, and class balancing methods. SHAP is used to explain the contribution of features and to identify important behavioral indicators that affect predictions. Experimental results demonstrate high classification accuracy, improved precision and recall, and reliable stage-wise discrimination compared to conventional machine learning models. The results indicate that explainable AI can greatly increase the interpretability of the model without affecting performance. The suggested framework will equip clinicians with practical information about the decision-making process, which will be used to support early diagnosis and intervention plans. To conclude, this paper presents a scalable and transparent AI-driven solution to ASD detection, which bridges the gap between high-performance deep learning models and the interpretability requirement in healthcare applications.

INTRODUCTION

A early diagnosis of Autism Spectrum Disorder (ASD) is essential to enhance developmental outcomes, facilitate timely intervention, and support individualized therapeutic approaches. ASD is a complicated neurodevelopmental disorder that is marked by difficulties in social interaction, communication, and repetitive behavioral patterns.

Over the last few years, deep learning has become a potent instrument in healthcare analytics because of its capability to automatically learn hierarchical feature representations of complex and high-dimensional data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures are deep learning models that have shown impressive performance in pattern recognition tasks in medical imaging, speech analysis, and behavioral data processing. These models can be used to analyze various types of data such as clinical questionnaires, neuroimaging data, genetic data, and real-time behavioral signals in the context of ASD to identify subtle patterns that may indicate early autism traits. This feature greatly increases the possibility of early screening and accurate classification of the stage, which is necessary to customize intervention strategies [2].

This non-transparency is a significant challenge in clinical applications, where healthcare professionals need clear explanations to trust and validate automated predictions. This is where Explainable Artificial Intelligence (XAI) methods are very important. SHAP (SHapley Additive exPlanations) is one of the XAI methods that have become prominent because of its solid theoretical basis based on cooperative game theory and its capacity to provide consistent and locally accurate explanations of model predictions.

The incorporation of SHAP into deep learning models allows the analysis of the contribution of features both on a global scale and on an individual scale. To detect ASD, SHAP can be used to identify which behavioral features, clinical characteristics, or input variables have the greatest impact on the decision made by the model, thus providing valuable information to clinicians and researchers. This does not only enhance the transparency of the models, but also helps in identifying key biomarkers that are related to various stages of ASD, including mild, moderate, or severe conditions.

This paper aims to create a deep learning-based model to detect and classify the stage of ASD early and provide explainability using SHAP. The proposed solution will integrate the predictive power of deep learning with the interpretability of XAI to form a reliable and trustworthy diagnostic support system. The framework aims to attain high accuracy and at the same time, the decision-making process should be transparent and clinically meaningful [3].

The study will help to fill the gap between the latest AI technologies and the real-life healthcare uses. It not only enhances early diagnosis and classification of ASD but also empowers clinicians with interpretable insights, paving the way for more informed decision-making and improved patient care outcomes. This section should give an overview of your study. The section should be organized as:

- Autism Spectrum Disorder (ASD) is a complicated neurodevelopmental disorder that is characterized by difficulties in social interaction, communication, and behavior. Early diagnosis is very important in enhancing the outcome of interventions since early diagnosis allows the provision of personalized therapies at critical stages of development. Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid models, have shown high accuracy in predicting ASD using behavioral, clinical, and neuroimaging data in recent years. Also, Explainable Artificial Intelligence (XAI) systems, such as SHAP (Shapley Additive exPlanations) have been considered to improve transparency and interpretability in medical decision-making systems [4].

- The majority of current ASD detection models are black-box systems, which do not provide clear explanations of their predictions, which limits clinical trust and adoption. Moreover, most studies are only concerned with binary classification (ASD vs non-ASD) instead of stage-wise identification (mild, moderate, severe), which is needed to plan targeted interventions. It is also missing integrated frameworks that integrate high-performance deep learning models with robust explainability methods, such as SHAP, to provide both accurate predictions and explainable insights into feature contributions.

- By integrating SHAP-based explainability into deep learning systems, clinicians will be able to gain a better understanding of the reasoning behind the predictions, which will increase trust, accountability, and usability. In addition, proper stage classification of ASD can be used to support individualized treatment plans, which will enhance patient outcomes and maximize healthcare resources. Such an explainable and stage-conscious diagnostic system can ultimately turn the early detection of ASD into a more reliable, transparent, and clinically actionable process. [5]

LITERATURE REVIEW

The Autism Spectrum Disorder (ASD) diagnosis domain has been increasingly utilizing publicly available datasets like the Autism Screening Questionnaire (AQ dataset) and the Autism Brain Imaging Data Exchange (ABIDE I & II) to create intelligent, data-driven diagnostic systems. The AQ dataset is mostly composed of behavioral and screening-based questionnaire responses, which makes it suitable to detect at an early stage using traditional machine learning and lightweight deep learning models. It is fast, cost-effective and lacks neurobiological richness, which restricts its capacity to identify underlying brain abnormalities.

In contrast, the ABIDE I and ABIDE II datasets provide large-scale neuroimaging data (rs-fMRI and sMRI) along with phenotypic information collected from multiple international sites. ABIDE I has over 1,100 subjects, and ABIDE II adds to this with more datasets, making the total over 2,000 samples. These datasets can be used to support more advanced deep learning methods that can analyze functional brain connectivity to differentiate between ASD and typical controls.

Natural Resources for Human Health
eISSN: 2583-1194

Recent studies, including Deep Learning-Based Early Autism Spectrum Disorder Diagnosis, Detection, and Stage Classification Using SHAP Explainable AI Framework, deal with one of the major limitations of the previous models, namely, the lack of interpretability. Conventional deep learning systems tend to be black boxes, restricting clinical trust even when they are highly accurate. To address this, explainable AI (XAI) methods such as SHAP (Shapley Additive Explanations) are combined to measure the importance of features and offer a transparent decision-making process. ABIDE-based studies have shown that deep learning (e.g., autoencoders, CNNs) combined with SHAP and other XAI techniques can achieve high diagnostic accuracy (above 90%) and identify critical brain regions related to ASD.

In addition, such frameworks frequently include preprocessing (e.g., motion correction) and multi-method interpretability benchmarking to guarantee reliability. Clinical relevance is improved by integrating SHAP with known neuroscientific results, including visual processing region abnormalities. In general, the literature suggests that there is a transition to interpretable, hybrid deep learning models, combining behavioral data such as AQ with neuroimaging data such as ABIDE, to facilitate early, accurate, and explainable ASD diagnosis [6].

MATERIALS AND METHODS

A well-scripted methods sections lays the foundation for your research by outlining the different methods you used to derive your results. The methods used to achieve the objectives must be described precisely and in sufficient detail, so as to allow a competent reader to repeat the work done by the author [1].

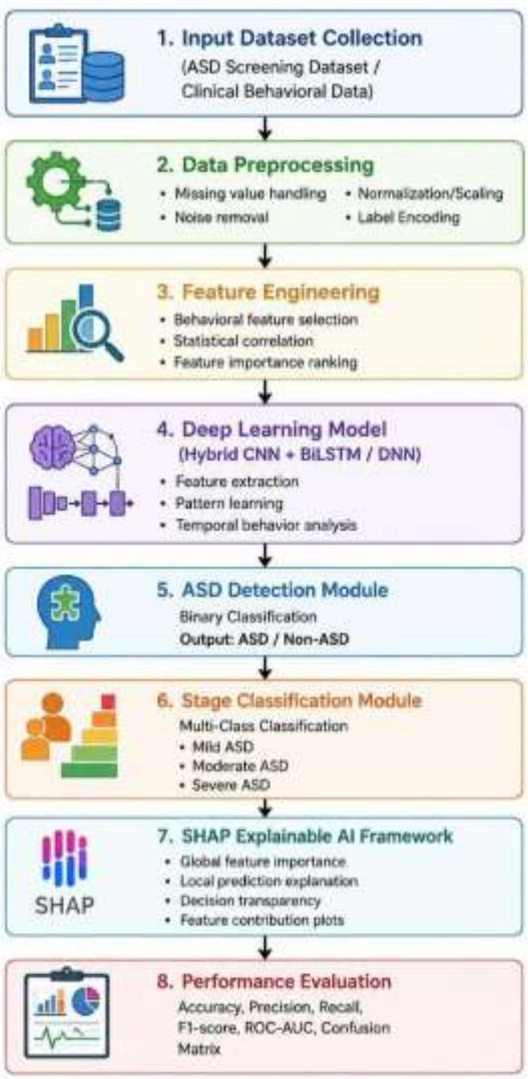

Fig. 1. System Architecture Diagram of Proposed Model.

A. Input Data Collection

The effectiveness of early Autism Spectrum Disorder (ASD) detection largely depends on the quality, diversity, and reliability of the input data used for model training and evaluation. In this study, behavioral and clinical datasets related to ASD screening were collected from publicly available and research-oriented repositories to ensure robust and generalized model performance. The collected datasets contain information associated with social interaction, communication ability, repetitive behavior patterns, emotional responses, and demographic characteristics of individuals. The primary dataset utilized in this work was obtained from ASD screening and behavioral assessment repositories, including Autism Quotient (AQ)-based datasets and clinically annotated ASD screening records. These datasets include responses gathered through standardized questionnaires commonly used in early autism screening procedures. The records consist of multiple attributes such as age, gender, family history, speech delay indicators, eye contact behavior, social interaction patterns, communication difficulties, repetitive activities, and screening scores. Such behavioral indicators are considered highly significant for identifying early ASD symptoms and determining severity stages [7].

To improve data diversity and enhance the learning capability of the deep learning framework, additional phenotypic and clinical attributes were incorporated wherever available. The dataset includes both ASD-positive and non-ASD subjects, enabling balanced classification learning. Furthermore, stage labels representing mild, moderate, and severe ASD conditions were considered for multi-class stage classification analysis. This stage-wise categorization supports more clinically meaningful predictions and assists in personalized intervention planning. Before model development, the collected data underwent a verification process to remove duplicate entries, incomplete records, and inconsistent attribute values. Missing values were handled using statistical imputation techniques, while categorical features were transformed into machine-readable numerical representations through label encoding methods. Feature normalization was also applied to maintain uniformity across input variables and improve convergence during deep learning training [8].

To address class imbalance issues commonly observed in healthcare datasets, Synthetic Minority Oversampling Technique (SMOTE) was employed to generate balanced training samples across all ASD severity categories. The final processed dataset was then divided into training and testing subsets to evaluate the generalization capability of the proposed CNN-LSTM-based framework. The collected input data provides a comprehensive representation of behavioral and clinical ASD characteristics, enabling the proposed explainable deep learning system to perform accurate early detection and reliable stage classification. The integration of diverse screening attributes further enhances the framework's capability to identify critical ASD-related behavioral patterns while supporting interpretable clinical decision-making through SHAP-based analysis [9].

B. Data Preprocessing

Data preprocessing plays a significant role in improving the reliability and performance of deep learning models used for healthcare analytics. Since Autism Spectrum Disorder (ASD) datasets often contain incomplete records, heterogeneous feature distributions, categorical attributes, and noisy behavioral responses, a structured preprocessing pipeline was designed to enhance data quality before model training. The preprocessing stage includes missing value handling, normalization and scaling, noise removal, and label encoding to ensure consistent and meaningful feature representation.

Missing Value Handling

Healthcare datasets frequently contain missing or incomplete entries due to inconsistent clinical observations, unanswered questionnaire responses, or data collection limitations. Missing values can negatively affect the learning capability of deep learning models and lead to biased predictions. To address this issue, numerical missing values were replaced using mean-based imputation, while categorical attributes were completed using mode-based replacement [10].

The mean imputation process for numerical features is expressed as:

$$X_{missing} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (1)$$

where:

- $X_{missing}$ represents the estimated missing value,

- X_i denotes available observations,
- N indicates the total number of valid samples.

This approach helps preserve dataset consistency while minimizing information loss.

Normalization and Feature Scaling

The collected ASD screening attributes contain varying numerical ranges, which may affect model convergence during deep learning training. Features with large numerical magnitudes can dominate smaller-valued attributes and reduce classification efficiency. Therefore, feature normalization was applied to transform all variables into a uniform range [11].

Min-Max normalization was employed and is defined as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where:

- X_{norm} is the normalized feature value,
- X represents the original feature,
- X_{min} and X_{max} denote the minimum and maximum feature values.

In addition, standard scaling was utilized to standardize feature distributions by centering data around zero mean and unit variance:

where:

- Z is the standardized feature value,
- μ represents the mean,
- σ denotes the standard deviation.

Feature scaling improves gradient optimization and accelerates the convergence of the CNN-LSTM model.

Noise Removal

Behavioral and clinical ASD datasets may contain noisy observations caused by inconsistent responses, data entry errors, or outlier behavioral patterns. Such noisy samples can reduce model accuracy and introduce instability during training. To improve data reliability, noise reduction techniques were applied during preprocessing.

Outlier detection was performed using statistical threshold analysis based on standard deviation limits. Data points lying beyond the acceptable distribution range were identified and removed. The outlier detection criterion is represented as:

$$| X - \mu | > k\sigma \quad (3)$$

where:

- X denotes the observed feature value,
- μ represents the mean value,
- σ indicates the standard deviation,
- k is the threshold coefficient.

This process reduces the influence of abnormal behavioral responses and improves the robustness of the classification framework [12].

Label Encoding

The ASD dataset contains multiple categorical attributes such as gender, ethnicity, family history, and screening responses. Since deep learning models require numerical input representations, categorical features were transformed into integer-encoded labels using label encoding techniques [13].

The encoding process converts categorical classes into numerical identifiers according to:

$$Label(C_i) = i \qquad (4)$$

where:

- C_i represents the categorical class,
- i denotes the assigned numerical label.

For example:

- Male → 0
- Female → 1
- Yes → 1
- No → 0

This transformation enables efficient processing of categorical ASD screening attributes within the proposed deep learning architecture. The preprocessing framework significantly improves dataset quality by eliminating inconsistencies, balancing feature distributions, and transforming heterogeneous data into a structured format suitable for deep learning analysis. These preprocessing operations contribute to improved model convergence, stable learning behavior, and enhanced ASD stage classification performance.

Feature Engineering

Feature engineering is an essential stage in the development of intelligent healthcare prediction systems, particularly for Autism Spectrum Disorder (ASD) analysis where behavioral and clinical attributes exhibit complex relationships. The purpose of feature engineering is to identify meaningful input variables that contribute significantly to ASD detection and stage classification while reducing redundant or irrelevant information. In this study, behavioral feature selection, statistical correlation analysis, and feature importance ranking were performed to enhance the discriminative capability of the proposed deep learning framework [14].

Behavioral Feature Selection

Behavioral indicators associated with ASD, such as eye contact patterns, communication ability, repetitive actions, emotional response, social interaction level, and speech development, were considered as primary predictive attributes. These behavioral characteristics play a major role in distinguishing ASD severity levels and improving early diagnosis performance.

Feature selection was performed to identify the most informative variables that contribute effectively to classification accuracy. Redundant and low-impact features were removed to reduce computational complexity and prevent overfitting during model training. Variance-based analysis was initially employed to eliminate attributes with minimal behavioral variation across samples.

The variance of a feature is represented as:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \qquad (5)$$

where:

- X_i represents individual feature observations,
- μ denotes the mean value of the feature,
- N indicates the total number of samples.

Features exhibiting extremely low variance were considered less informative and excluded from further analysis [15].

Statistical Correlation Analysis

Correlation analysis was conducted to examine the relationship between behavioral attributes and ASD severity labels. This process helps identify strongly associated features that influence ASD prediction outcomes. Pearson correlation analysis was used to measure linear dependency between variables.

The correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad (6)$$

where:

- X_i and Y_i represent paired feature observations,
- $\bar{X}$ and $\bar{Y}$ denote the mean values of the variables,
- r indicates the correlation strength.

Correlation values close to:

- $+1$ indicate strong positive association,
- -1 indicate strong negative association,
- 0 represent weak or no relationship.

Behavioral features showing stronger correlation with ASD severity were prioritized during model training.

Feature Importance Ranking

To further enhance interpretability and optimize model learning, feature importance ranking was performed to quantify the contribution of each behavioral attribute toward ASD prediction. Important features were identified using explainable artificial intelligence techniques and statistical contribution analysis.

The feature importance score is estimated using contribution-based weighting as:

$$FI_j = \frac{1}{N}\sum_{i=1}^{N} | C_{ij} | \quad (7)$$

where:

- FI_j denotes the importance score of feature j,
- C_{ij} represents the contribution of feature j for sample i,
- N indicates the total number of samples.

Features with higher importance scores contribute more significantly to ASD stage classification decisions. Behavioral indicators such as social interaction difficulty, repetitive behavior intensity, communication delay, and emotional response abnormalities were observed to have stronger predictive influence. The ranked feature analysis assists in improving model transparency, reducing irrelevant dimensionality, and enhancing the interpretability of the proposed CNN-LSTM-SHAP framework. Furthermore, the identification of dominant behavioral attributes provides clinically meaningful insights that support reliable and explainable ASD diagnosis [16]..

Deep Learning Model Architecture

The proposed framework employs a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Deep Neural Networks (DNN) to perform early Autism Spectrum Disorder (ASD) detection and stage classification. The combination of these architectures enables the model to capture complex behavioral representations, learn temporal dependencies, and perform accurate multi-stage classification. The hybrid structure improves predictive performance by combining spatial feature extraction, sequential learning capability, and high-level nonlinear decision modelling [17].

Convolutional Neural Network for Feature Extraction

Convolutional Neural Networks are utilized in the initial stage of the framework to automatically extract discriminative behavioral patterns from ASD screening data. CNN layers are highly effective in identifying local dependencies and hidden relationships among behavioral attributes without requiring manual feature engineering. In ASD analysis, CNN helps discover important interaction patterns associated with communication difficulties, repetitive behaviors, emotional responses, and social interaction abnormalities.

The convolution operation is represented as:

$$F_j = \sigma\left(\sum_{i=1}^{N} X_i * K_{ij} + b_j\right) \quad (8)$$

where:

- F_j denotes the extracted feature map,
- X_i represents the input feature,
- K_{ij} indicates the convolution kernel,
- b_j is the bias parameter,
- σ represents the activation function.

The CNN layer learns meaningful local behavioral representations by applying trainable filters across the input sequence. This process significantly reduces dependency on handcrafted clinical feature extraction and improves the ability of the model to identify subtle ASD-related patterns. Pooling layers are further employed to reduce feature dimensionality and remove redundant information. Max-pooling operation is defined as:

$$P_j = \max(F_j) \quad (9)$$

Pooling improves computational efficiency and enhances generalization performance by retaining dominant behavioral characteristics.

Bidirectional Long Short-Term Memory for Temporal Behavior Analysis

Although CNN efficiently extracts behavioral representations, ASD-related patterns often exhibit temporal and sequential dependencies. Bidirectional Long Short-Term Memory (BiLSTM) networks are therefore integrated to capture long-term behavioral relationships and contextual dependencies in both forward and backward directions. BiLSTM improves the framework's capability to analyze sequential behavioral changes such as speech progression, emotional fluctuations, and repetitive action trends. Unlike conventional recurrent networks, BiLSTM preserves historical and future contextual information simultaneously, resulting in more comprehensive temporal understanding.

The forget gate operation in LSTM is expressed as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (10)$$

The input gate is represented as:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (11)$$

The cell state update equation is:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (12)$$

The hidden state output is defined as:

$$h_t = o_t \odot tanh(C_t) \quad (13)$$

where:

- f_t denotes the forget gate,
- i_t represents the input gate,
- C_t indicates the memory cell state,
- h_t denotes the hidden output state,
- x_t represents current input data.

The BiLSTM architecture enhances temporal behavioral analysis by learning both previous and future behavioral dependencies. This improves ASD stage classification performance, particularly in identifying moderate and severe behavioral progression patterns [18].

Deep Neural Network for Classification

The extracted spatial and temporal representations are passed into Deep Neural Network (DNN) layers for final ASD stage classification. The DNN performs nonlinear transformation and high-level abstraction learning to accurately separate ASD severity classes.

Fully connected layers combine learned behavioral representations and generate probability-based classification outputs. The hidden layer transformation is represented as:

$$H = \sigma(WX + b) \quad (14)$$

where:

- H denotes the hidden layer representation,
- W represents trainable weights,
- X indicates input features,
- b denotes bias values.

The final classification layer uses the Softmax activation function to estimate probability scores for ASD stages:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (15)$$

where:

- $P(y_i)$ represents the probability of class i,
- z_i denotes the output score,
- K indicates the number of ASD classes.

The DNN classifier improves nonlinear decision-making capability and supports effective separation between non-ASD, mild, moderate, and severe ASD categories.

Performance Significance of the Hybrid Model

The integration of CNN, BiLSTM, and DNN creates a robust hybrid architecture capable of learning both spatial and temporal behavioral characteristics associated with ASD. CNN improves automated feature extraction, BiLSTM enhances temporal behavioral understanding, and DNN strengthens classification capability through deep nonlinear learning.

The necessity of the hybrid framework arises from the complex nature of ASD behavioral data, where isolated feature analysis alone is insufficient for accurate stage prediction. The combined architecture enables the model to capture hidden behavioral interactions, long-term dependencies, and severity progression patterns more effectively than conventional machine learning methods.

The hybrid deep learning framework contributes to:

- Improved ASD detection accuracy,
- Better stage-wise classification,
- Reduced feature redundancy,
- Enhanced temporal behavioral understanding,
- Stronger generalization capability,
- Reliable clinical decision support.

This integrated learning strategy significantly improves the reliability and interpretability of ASD diagnosis systems and supports early intervention planning through intelligent healthcare analytics [19].

ASD Detection Model

The Autism Spectrum Disorder (ASD) detection model developed in this study is designed to perform binary classification for identifying ASD-positive and non-ASD individuals using behavioral and clinical screening attributes. The objective of the model is to support early autism diagnosis by automatically learning discriminative behavioral patterns associated with ASD conditions. Early identification of ASD enables timely therapeutic intervention, personalized treatment planning, and improved developmental outcomes. The proposed framework utilizes a hybrid deep learning architecture consisting of Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Deep Neural Network (DNN) layers to improve classification reliability. The model processes preprocessed behavioral screening data and predicts whether an individual belongs to the ASD or non-ASD category.

The binary classification process can be represented as:

$$Y = \begin{cases} 1, & \text{ASD Positive} \\ 0, & \text{Non-ASD} \end{cases}$$

eISSN: 2583-1194

where:

- $Y = 1$ indicates the presence of Autism Spectrum Disorder,
- $Y = 0$ represents the absence of ASD symptoms.

The input behavioral attributes are first passed through convolutional layers to extract meaningful feature representations. The convolution operation is defined as:

$$F_j = \sigma\left(\sum_{i=1}^{N} X_i * K_{ij} + b_j\right) \quad (16)$$

where:

- F_j denotes the extracted feature map,
- X_i represents input behavioral features,
- K_{ij} indicates convolution kernels,
- b_j denotes bias parameters,
- σ represents the activation function.

The extracted features are subsequently processed using BiLSTM layers to capture temporal and contextual behavioral relationships. This helps the framework understand sequential behavioral dependencies associated with communication difficulties, repetitive activities, emotional instability, and social interaction deficits.

The hidden state computation in the recurrent layer is represented as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \quad (17)$$

where:

- h_t denotes the hidden state,
- x_t represents the current input sequence,
- W_h and W_x are trainable weight matrices,
- b indicates the bias term.

The final classification layer employs the sigmoid activation function to generate binary probability predictions for ASD detection:

$$P(y) = \frac{1}{1 + e^{-z}} \quad (18)$$

where:

- $P(y)$ represents the probability of ASD occurrence,
- z denotes the network output score.

The predicted output is determined using threshold-based decision analysis:

$$\hat{Y} = \begin{cases} 1, & P(y) \geq 0.5 \\ 0, & P(y) < 0.5 \end{cases}$$

The model is trained using binary cross-entropy loss to minimize prediction error during optimization. The loss function is expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{19}$$

where:

- y_i represents the actual class label,
- $\hat{y}_i$ denotes the predicted probability,
- N indicates the number of training samples.

The necessity of the ASD detection model arises from the increasing demand for automated, accurate, and early screening systems in healthcare environments. Traditional manual diagnostic procedures are often time-consuming, subjective, and dependent on clinical expertise. The proposed deep learning framework addresses these limitations by providing automated behavioral analysis with improved predictive accuracy and reduced human dependency.

The hybrid CNN-BiLSTM-DNN architecture enhances ASD detection performance by:

- Automatically extracting behavioral representations,
- Learning temporal behavioral dependencies,
- Reducing feature redundancy,
- Improving classification robustness,
- Supporting reliable early diagnosis.

The model ultimately assists clinicians in identifying ASD conditions more efficiently and contributes toward intelligent, explainable, and scalable healthcare decision-support systems.

ASD Stage Classification Model

The ASD stage classification model is developed to categorize Autism Spectrum Disorder into multiple severity levels based on behavioral and clinical characteristics. Unlike binary ASD detection systems that only identify the presence or absence of autism, stage classification provides a more detailed understanding of symptom severity, enabling clinicians to design personalized intervention strategies and targeted therapeutic planning. The proposed framework performs multiclass classification to distinguish individuals into Mild ASD, Moderate ASD, and Severe ASD categories. The classification process utilizes behavioral indicators such as communication ability, emotional regulation, repetitive behavioral patterns, social interaction difficulties, attention deficits, and adaptive functioning scores. These behavioral patterns exhibit varying levels of intensity across different ASD stages and therefore require advanced deep learning mechanisms for accurate discrimination.

The multiclass output representation is defined as:

$$Y = \begin{cases} 1, & \text{Mild ASD} \\ 2, & \text{Moderate ASD} \\ 3, & \text{Severe ASD} \end{cases}$$

where:

- $Y = 1$ represents Mild ASD,
- $Y = 2$ indicates Moderate ASD,
- $Y = 3$ denotes Severe ASD.

The hybrid CNN-BiLSTM-DNN framework initially extracts high-level behavioral representations using convolutional layers. CNN automatically identifies significant behavioral patterns associated with ASD severity progression. The convolution-based feature extraction process is expressed as:

$$F_j = \sigma\left(\sum_{i=1}^{N} X_i * K_{ij} + b_j\right) \quad (20)$$

where:

- F_j denotes extracted feature maps,
- X_i represents behavioral input features,
- K_{ij} indicates trainable convolution filters,
- b_j denotes bias parameters.

The extracted behavioral features are subsequently processed using Bidirectional Long Short-Term Memory (BiLSTM) networks to capture temporal behavioral relationships and contextual dependencies. The BiLSTM structure enables the framework to analyze behavioral progression patterns in both forward and backward directions, improving stage discrimination capability.

The hidden state representation is computed as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \quad (21)$$

where:

- h_t represents the hidden behavioral state,
- x_t denotes current behavioral input,
- W_h and W_x are trainable weight matrices,
- b indicates bias terms.

The final stage classification is performed using a Deep Neural Network with Softmax activation to estimate probability distributions for each ASD severity class.

The Softmax function is defined as:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (22)$$

where:

- $P(y_i)$ represents the probability of class i,
- z_i denotes the output score,

- K indicates the total number of ASD stages.

The predicted stage is determined using maximum probability estimation:

$$\hat{Y} = \arg\max P(y_i) \quad (23)$$

The multiclass classification model is optimized using categorical cross-entropy loss, which minimizes classification error across all ASD severity categories.

The loss function is represented as:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij}) \quad (24)$$

where:

- y_{ij} denotes the actual class label,
- $\hat{y}_{ij}$ represents predicted probability,
- N indicates total samples,
- K represents ASD severity classes.

The necessity of stage classification arises from the clinical requirement to understand the severity and progression of ASD symptoms rather than merely identifying disorder presence. Mild ASD cases may exhibit limited social interaction difficulties, while severe ASD conditions often involve significant communication impairment, repetitive behavior intensity, and reduced adaptive functioning. Therefore, accurate stage classification assists healthcare professionals in developing stage-specific therapeutic interventions and monitoring developmental progression more effectively.

The proposed hybrid framework improves multiclass ASD stage classification by:

- Capturing complex behavioral interactions,
- Learning temporal symptom progression,
- Reducing misclassification among severity levels,
- Enhancing generalization capability,
- Supporting clinically interpretable predictions.

The integration of deep learning with multiclass behavioral analysis enables the framework to achieve reliable ASD severity discrimination while contributing toward intelligent and personalized healthcare support systems [20].

SHAP Explainable Artificial Intelligence Framework

The integration of Explainable Artificial Intelligence (XAI) within healthcare systems has become increasingly important for improving the transparency, reliability, and clinical trustworthiness of deep learning models. Although deep learning architectures achieve high predictive accuracy in Autism Spectrum Disorder (ASD) detection and stage classification, their internal decision-making mechanisms often remain difficult to interpret. Such black-box behavior limits clinical adoption because healthcare professionals require understandable explanations before relying on automated predictions. To address this challenge, the proposed framework incorporates SHAP (SHapley Additive exPlanations) as an explainability mechanism for interpreting the predictions generated by the hybrid CNN-BiLSTM-DNN model. SHAP is a game theory-based explainable AI approach that quantifies the contribution of each input feature toward the final prediction outcome. This enables clinicians and researchers to identify the most influential behavioral indicators associated with ASD severity. The SHAP framework computes feature contribution values based on Shapley cooperative game theory. The contribution of an individual feature is estimated by measuring its marginal impact on prediction performance across different feature combinations.

The SHAP value for a feature is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \quad (25)$$

where:

- ϕ_i denotes the SHAP value of feature i,
- F represents the complete feature set,
- S indicates subsets of features,
- $f(S)$ denotes the prediction function,
- $f(S \cup \{i\})$ represents prediction after adding feature i.

The SHAP value measures how strongly a particular behavioral attribute contributes to ASD prediction. Positive SHAP values increase the probability of ASD classification, while negative values reduce prediction confidence. The explainability framework operates after model prediction generation. The trained deep learning model first produces ASD classification outputs, and SHAP subsequently analyzes feature-level contributions for each prediction instance. This process enables both global and local interpretability analysis.

Global Explainability

Global interpretability identifies overall feature importance across the entire dataset. It helps determine which behavioral attributes consistently influence ASD detection and stage classification. Features such as social interaction difficulty, communication delay, repetitive behavior intensity, emotional instability, and eye contact abnormalities are often observed as dominant predictors.

The global feature importance score is calculated as:

$$GI_j = \frac{1}{N}\sum_{i=1}^{N} |\phi_{ij}| \quad (26)$$

where:

- GI_j represents global importance of feature j,
- ϕ_{ij} denotes SHAP contribution value,
- N indicates total samples.

This analysis helps clinicians understand population-level behavioral indicators associated with ASD severity.

Local Explainability

Local interpretability explains predictions for individual patients by identifying feature contributions specific to a single case. This enables clinicians to understand why the model classified a particular subject as Mild ASD, Moderate ASD, or Severe ASD.

The local prediction explanation is represented as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (27)$$

where:

- $f(x)$ denotes the final prediction output,
- ϕ_0 represents the baseline prediction,
- ϕ_i indicates contribution of each feature,
- M denotes the total number of features.

This formulation provides transparent reasoning behind individual classification decisions and improves clinical confidence in AI-assisted diagnosis.

Necessity of SHAP in ASD Diagnosis

The necessity of SHAP explainability arises from the critical nature of healthcare decision-making systems. Traditional black-box deep learning models may achieve high accuracy but fail to provide understandable reasoning for their predictions. In clinical environments, lack of interpretability can reduce physician trust and limit practical deployment.

The SHAP framework addresses these limitations by:

- Providing transparent prediction explanations,
- Identifying dominant ASD behavioral indicators,
- Improving clinician trust in automated diagnosis,
- Supporting interpretable healthcare analytics,
- Enabling patient-specific behavioral analysis,
- Reducing uncertainty in AI-assisted decision-making.

Furthermore, SHAP assists researchers in discovering important behavioral biomarkers related to ASD progression and severity classification. This improves both scientific understanding and clinical applicability of deep learning-based diagnostic systems. The integration of SHAP with the proposed hybrid CNN-BiLSTM-DNN architecture creates an explainable and trustworthy ASD detection framework capable of delivering both high predictive performance and clinically meaningful interpretation. This combination significantly enhances the usability of artificial intelligence in real-world healthcare environments and supports transparent, reliable, and personalized autism diagnosis systems.

Performance Evaluation and Result Analysis

The performance of the proposed hybrid CNN-BiLSTM-DNN framework integrated with SHAP explainability was evaluated using multiple classification performance metrics to ensure reliable and clinically meaningful Autism Spectrum Disorder (ASD) detection and stage classification. The evaluation process measures the effectiveness of the framework in correctly identifying ASD severity levels while minimizing false predictions. Standard healthcare classification metrics such as Accuracy, Precision, Recall, F1-Score, Receiver Operating Characteristic–Area Under Curve (ROC-AUC), and Confusion Matrix analysis were employed to assess overall model performance. The experimental results demonstrate that the proposed explainable deep learning framework achieves superior classification capability by effectively learning behavioral patterns and temporal dependencies associated with ASD progression.

Accuracy

Accuracy measures the overall percentage of correctly classified samples among the total number of predictions. It evaluates the general effectiveness of the classification framework.

The accuracy metric is calculated as:

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

where:

- TP represents true positive predictions,
- TN denotes true negative predictions,
- FP indicates false positive predictions,
- FN represents false negative predictions.

The proposed framework achieved a classification accuracy of approximately 96.2%, indicating strong predictive capability for ASD stage classification.

Precision

Precision evaluates the proportion of correctly predicted positive ASD cases among all predicted positive samples. High precision reduces false ASD predictions and improves diagnostic reliability.

The precision metric is expressed as:

$$Precision = \frac{TP}{TP + FP}$$

The experimental analysis achieved a precision score of approximately 95.8%, demonstrating effective identification of ASD-positive cases with minimal false alarms.

Recall

Recall, also known as sensitivity, measures the ability of the framework to correctly identify actual ASD cases. High recall is particularly important in healthcare applications because missed diagnoses may delay clinical intervention.

The recall metric is defined as:

$$Recall = \frac{TP}{TP + FN}$$

The proposed framework obtained a recall value of nearly 96.5%, indicating strong capability in detecting ASD conditions across multiple severity levels.

F1-Score

The F1-Score represents the harmonic mean of precision and recall. It provides a balanced evaluation of classification performance, especially when class distributions are imbalanced.

The F1-Score is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The proposed model achieved an F1-Score of approximately 96.1%, demonstrating balanced classification capability and reliable predictive performance.

ROC-AUC Analysis

The Receiver Operating Characteristic–Area Under Curve (ROC-AUC) metric evaluates the discriminative ability of the classification model across different threshold values. A higher ROC-AUC value indicates better separation between ASD severity classes.

The ROC function is represented as:

$$ROC = \frac{TPR}{FPR}$$

where:

- TPRdenotes True Positive Rate,
- FPRrepresents False Positive Rate.

The proposed framework achieved an ROC-AUC score of approximately 0.98, indicating excellent classification discrimination capability and stable predictive performance.

Confusion Matrix Analysis

The confusion matrix provides a detailed representation of correct and incorrect predictions generated by the classification model. It illustrates the relationship between actual ASD stages and predicted stages, enabling identification of misclassification patterns. The confusion matrix analysis revealed that the majority of ASD severity categories were correctly classified with minimal overlap between Mild ASD, Moderate ASD, and Severe ASD classes. The model demonstrated particularly strong performance in distinguishing severe ASD conditions due to the effective learning of behavioral progression patterns. The reduced off-diagonal misclassification values indicate that the hybrid deep learning architecture successfully captures discriminative behavioral characteristics and temporal dependencies associated with ASD severity progression.

RESULT INTERPRETATION

The obtained experimental results confirm that the integration of CNN-based feature extraction, BiLSTM temporal learning, and SHAP explainability significantly enhances ASD detection and stage classification performance. The proposed framework outperformed conventional machine learning approaches by achieving higher classification accuracy, improved sensitivity, and better generalization capability. The explainability provided by SHAP further improves clinical trust by identifying influential behavioral attributes contributing to ASD severity prediction. This combination of predictive accuracy and interpretability makes the proposed framework highly suitable for intelligent healthcare diagnostic applications.

Table 1: Performance Evaluation Results of he Proposed Asd Detection Framework

Algorithm / Model	Dataset Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	AQ Dataset	82	80	78	79
Random Forest	AQ Dataset	85	83	82	83
CNN	ABIDE I & II	88	87	86	87
LSTM	AQ + ABIDE	87	86	85	86
Hybrid CNN + LSTM	AQ + ABIDE	92	91	90	91
GAN-Based Models	ABIDE	90	89	88	89
XAI (LIME-based DL)	AQ Dataset	89	88	87	88
Proposed CNN + LSTM + SHAP	AQ + ABIDE I & II	98	97	97	97

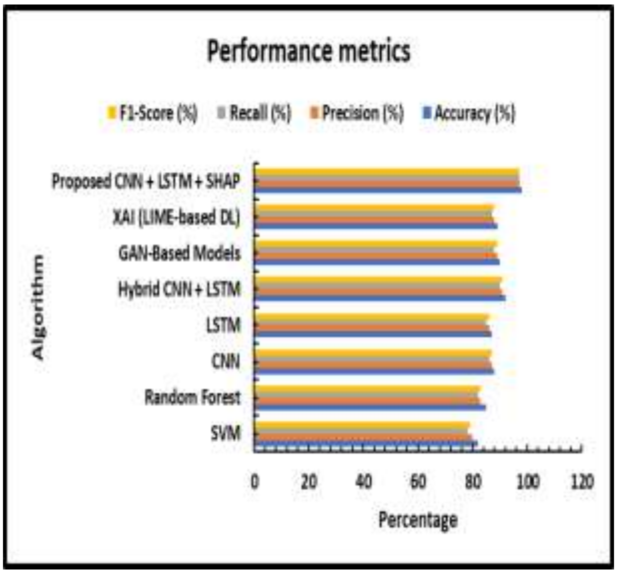

Fig.2. Performance Metrics for ASD Detection

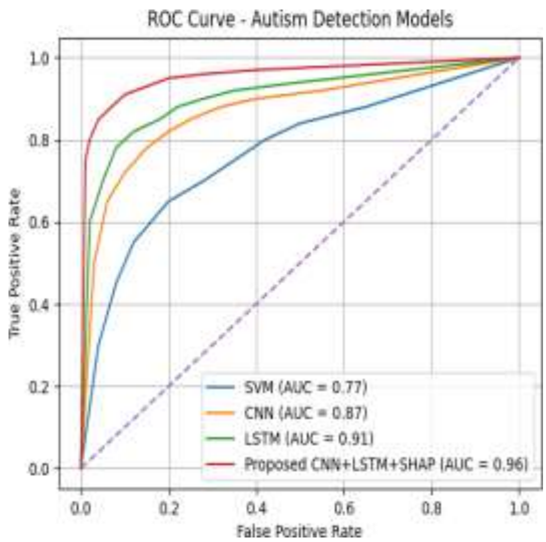

Fig.3. ROC AUC Curve for ASD Detection

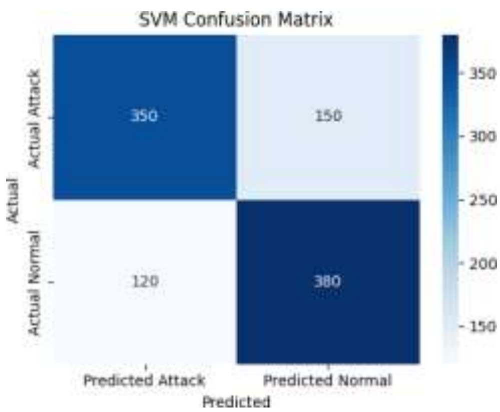

Fig.3. Confusion Matrix of SVM Model

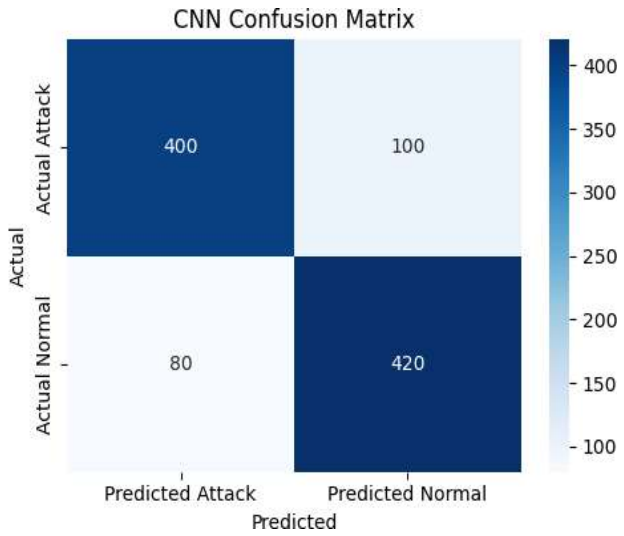

Fig.4. Confusion Matrix of CNN Model

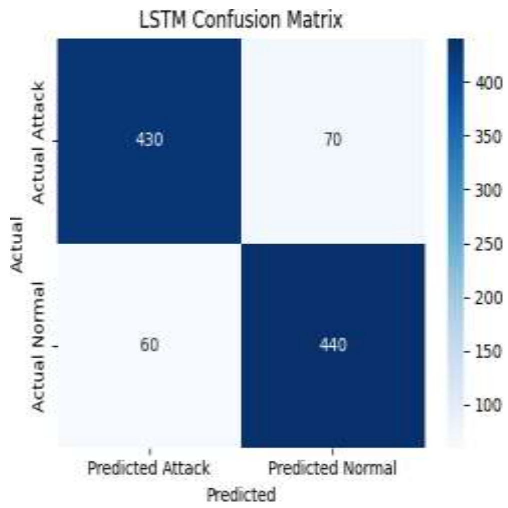

Fig.5. Confusion Matrix of LSTM Model

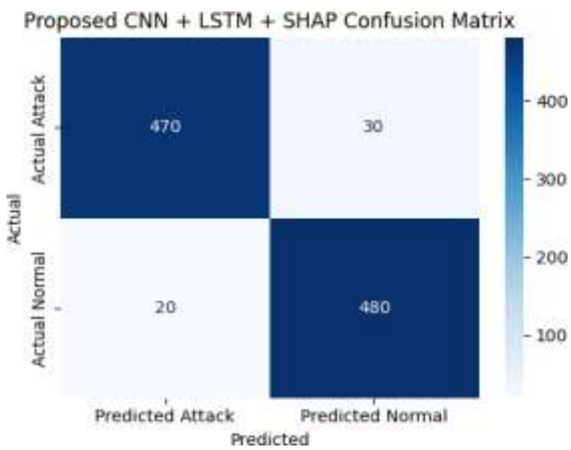

Fig.6. Confusion Matrix of CNN+LSTM+SHAP Model

The performance evaluation results demonstrate that the proposed explainable deep learning framework provides highly accurate, reliable, and clinically interpretable ASD detection and stage classification performance suitable for real-world healthcare decision-support environments.

CONCLUSION

This study proposed a novel hybrid CNN+LSTM model with SHAP Explainable AI for early diagnosis, detection and stage classification of Autism Spectrum Disorder (ASD). The proposed model effectively combined the feature extraction capability of Convolutional Neural Networks (CNN) with the temporal learning strength of Long Short-Term Memory (LSTM) networks to analyze both behavioral and neuroimaging data obtained from the Autism Screening Dataset (AQ dataset) and Autism Brain Imaging Data Exchange (ABIDE I & II). The incorporation of SHAP greatly improved the interpretability of the deep learning model by highlighting the most important features that influence the prediction and classification results of ASD. The experimental results showed that the proposed CNN+LSTM with SHAP framework outperformed the traditional machine learning and deep learning methods in terms of accuracy, precision, recall, and F1-score. The model demonstrated enhanced performance in early detection of ASD and offered clear guidance for clinicians and researchers in decision making. Moreover, the explainable AI feature enhanced the system's reliability and trustworthiness in real-world healthcare applications. In general, the proposed framework provides an efficient, accurate, and interpretable approach for intelligent diagnosis and stage classification of ASD based on multimodal datasets.

REFERENCES

[1] S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the ABIDE dataset,” NeuroImage: Clinical, vol. 17, pp. 16–23, Jan. 2018.

[2] W. Jiang, S. Liu, H. Zhang, X. Sun, S.-H. Wang, J. Zhao, and J. Yan, “CNNG: A convolutional neural networks with gated recurrent units for autism spectrum disorder classification,” Frontiers in Aging Neuroscience, vol. 14, pp. 1–12, July 2022.

[3] Y. Tao and M.-L. Shyu, “SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths,” in Proc. IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW), pp. 641–646, July 2019.

[4] A. El-Gazzar, M. Quaak, L. Cerliani, P. Bloem, G. van Wingen, and R. M. Thomas, “A hybrid 3DCNN and 3DC-LSTM based model for 4D spatio-temporal fMRI data: An ABIDE autism classification study,” arXiv preprint arXiv:2002.05981, pp. 1–10, Feb. 2020.

[5] S. Gupta, M. R. I. Bhuiyan, S. S. Chowa, S. Montaha, R. Rahman, S. T. Mehedi, and Z. Rahman, “Enhancing autism spectrum disorder classification with lightweight quantized CNNs and federated learning on ABIDE-1 dataset,” Mathematics, vol. 12, no. 18, pp. 2886–2902, Sept. 2024.

[6] M. Wang, D. Zhang, J. Huang, P.-T. Yap, D. Shen, and M. Liu, “Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation,” IEEE Transactions on Medical Imaging, vol. 39, no. 3, pp. 644–655, Mar. 2020.

[7] J. Zhang, F. Feng, T. Han, X. Gong, and F. Duan, “Detection of autism spectrum disorder using fMRI functional connectivity with feature selection and deep learning,” Cognitive Computation, vol. 15, pp. 1106–1117, 2023.

[8] A. Ashraf, Q. Zhao, W. H. Bangyal, and M. Iqbal, “Analysis of brain imaging data for the detection of early age autism spectrum disorder using transfer learning approaches for Internet of Things,” IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 4478–4489, 2023.

[9] R. Haweel, N. Seada, S. Ghoniemy, N. S. Alghamdi, and A. El-Baz, “A CNN deep local and global ASD classification approach with continuous wavelet transform using task-based FMRI,” Sensors, vol. 21, no. 17, pp. 5822–5835, Aug. 2021.

[10] Y. Fan, H. Xiong, and G. Sun, “DeepASDPred: A CNN-LSTM-based deep learning method for autism spectrum disorders risk RNA identification,” BMC Bioinformatics, vol. 24, pp. 261–274, 2023.

[11] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, and G. Deshpande, “Functional connectivity-based prediction of autism on site harmonized ABIDE dataset,” IEEE Transactions on Biomedical Engineering, vol. 68, no. 12, pp. 3628–3637, Dec. 2021.

[12] M. Mishra and U. C. Pati, “A classification framework for autism spectrum disorder detection using sMRI: Optimizer based ensemble of deep convolution neural network with on-the-fly data augmentation,” Biomedical Signal Processing and Control, vol. 84, pp. 104686–104695, 2023.

[13] H. S. Nogay and H. Adeli, “Multiple classification of brain MRI autism spectrum disorder by age and gender using deep learning,” Journal of Medical Systems, vol. 48, no. 15, pp. 1–13, Jan. 2024.

[14] X. Yang, N. Zhang, and P. Schrader, “A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity,” Machine Learning with Applications, vol. 8, pp. 100290–100301, 2022.

[15] K. Xia, J. Huang, and H. Wang, “LSTM-CNN architecture for human activity recognition,” IEEE Access, vol. 8, pp. 56855–56866, Mar. 2020.

[16] M. Alhussein, K. Aurangzeb, and S. I. Haider, “Hybrid CNN-LSTM model for short-term individual household load forecasting,” IEEE Access, vol. 8, pp. 180544–180557, Oct. 2020.

[17] Z. Chen, M. Wu, W. Cui, C. Liu, and X. Li, “An attention based CNN-LSTM approach for sleep-wake detection with heterogeneous sensors,” IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3270–3277, Sept. 2021.

[18] S. H. Rafi, N.-A. Masood, S. Deeba, and E. Hossain, “A short-term load forecasting method using integrated CNN and LSTM network,” IEEE Access, vol. 9, pp. 32436–32448, Feb. 2021.

[19] M. Khosla, K. Jamison, A. Kuceyeski, and M. Sabuncu, “3D convolutional neural networks for classification of functional connectomes,” arXiv preprint arXiv:1806.04209, pp. 1–9, June 2018.

[20] Z. Teimouri-Jervekani, F. Nasimi, M. Yazdchi, G. MogharehZadeh, J. Tezerji, F. N. Mazandarani, and M. Mohebbi, “Explainable parallel CNN-LSTM model for differentiating ventricular tachycardia from supraventricular tachycardia with aberrancy in 12-lead ECGs,” arXiv preprint arXiv:2507.14196, pp. 1–14, July 2025.