ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2024.0481



# HYBRID TRANSFORMER-CNN MODELS FOR ENHANCED AUTISM SPECTRUM DISORDER CLASSIFICATION USING CLINICAL AND NEUROIMAGING DATA

**Sanju S Anand and Shashidhar Kini**
*Institute of Computer Science and Information Science, Srinivas University, India*

***Abstract***

***Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by a highly heterogeneous presentation, posing significant challenges for early diagnosis. The subtle differences between ASD and non-ASD individuals, especially during early developmental stages, make accurate classification difficult. However, early detection plays a crucial role in improving developmental outcomes through timely intervention, enabling affected children and families to access specialized therapies and support systems. This study explores the potential of using clinical data combined with deep learning techniques for automated ASD classification. We evaluated various deep learning models, including 3D CNN ResNet50, sequential CNN, 2D CNN combined with XGBoost, 2D CNN ResNet101, and Transformer-based architectures like the standard Transformer and Swin Transformer integrated with CNN. The incorporation of clinical parameters alongside neuroimaging features facilitated more nuanced pattern recognition associated with ASD. Conventional CNN models yielded moderate classification accuracy, ranging from 60% to 78%. Transformer-based models demonstrated superior performance, with Swin Transformer achieving the accuracy of 75%, highlighting their importance in capturing intricate patterns and relationships in the data. The Swin Transformer, or "Shifted Window Transformer," is a type of Vision Transformer (ViT) architecture designed for computer vision tasks. It introduces a hierarchical structure with multi-scale feature representation, making it more efficient for image recognition tasks compared to traditional ViTs. The results show that hybrid models, specifically the Hybrid CNN+Swin Transformer, outperform both traditional CNN architectures and pure transformer-based methods, achieving the maximum classification accuracy at 80%. This implies that a more thorough method of identifying ASD-related patterns in brain imaging data can be achieved by fusing the global contextual understanding of the Swin Transformer with CNN's spatial feature extraction capabilities. These findings underscore the potential of using Transformer-based architectures in ASD classification, leveraging clinical data to improve precision in early detection. This research provides a foundation for future investigations into hybrid approaches that integrate multiple data sources, advancing automated diagnostic systems for neurodevelopmental disorders.***

***Keywords:***

***Autism Spectrum Disorder (ASD), Deep Learning, Convolutional Neural Networks (CNN), Hybrid CNN-RNN Models, XGboost (Extreme Gradient Boosting), Neuroimaging, ASD Classification, Early Detection, Feature Extraction, Brain Imaging, 2D and 3D CNN, RNN (Recurrent Neural Networks), ASD Diagnosis, (aMRI)Anatomical MRI, Swin Transformer (Shifted Window Transformer)***

## 1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by a range of symptoms, including social communication deficits and repetitive behaviours. Early and accurate diagnosis is crucial for timely intervention, yet traditional diagnostic approaches based on behavioural assessments can be subjective and may not always capture the disorder's heterogeneity. With advancements in neuroimaging techniques such as anatomical Magnetic Resonance Imaging (aMRI) and machine learning, there is potential to enhance the objectivity and accuracy of ASD diagnosis. ASD typically manifests early in life, with challenges in social interaction, communication, and adaptive behaviors, often accompanied by repetitive or restrictive activities. Early identification and intervention are pivotal, as they can significantly influence developmental trajectories and improve quality of life. However, traditional diagnostic methods, relying primarily on behavioral observations and caregiver-reported histories, are inherently subjective and may overlook subtle neurological underpinnings. Advancements in neuroimaging have opened new pathways for understanding the structural and functional abnormalities associated with ASD. Techniques like anatomical Magnetic Resonance Imaging (aMRI) provide non-invasive means to capture high-resolution brain data, offering insights into potential biomarkers of the disorder. When combined with machine learning (ML), these datasets enable the exploration of intricate patterns that are challenging to discern through conventional analysis.

This research explores the integration of neuroimaging data and clinical assessments using deep learning models to classify ASD. We utilize 2D slices extracted from aMRI scans, combined with clinical data such as ADI-R and ADOS scores, to train machine learning models capable of discerning ASD-related patterns in brain structure. Our approach evaluates multiple model architectures, including convolutional neural networks (CNNs) and advanced transformer-based networks.

The methods tested in this study cover a range of deep learning techniques, including 3D CNN, ResNet50, Normal CNN (Sequential), 2D CNN combined with XGBoost, 2D CNN ResNet101, Transformer-based models, and the Swin Transformer. Each model leverages different aspects of neuroimaging data to enhance diagnostic accuracy. Performance evaluation across these models shows varying levels of classification accuracy, with results demonstrating 62% accuracy for the 3D CNN, 69% for ResNet50, 78% for the Normal CNN, 60% for 2D CNN + XGBoost, 60% for 2D CNN ResNet101, 70% for the Transformer, and 75% for the Swin Transformer. The findings show that hybrid models, particularly the Hybrid CNN+Swin Transformer, obtain the maximum classification accuracy of 80%, outperforming both traditional CNN architectures and pure transformer-based techniques. This shows that integrating CNN's spatial feature extraction capabilities with the Swin Transformer's global contextual awareness provides a more complete method for detecting ASD-related abnormalities in brain imaging data.

The proposed approach aims to capitalize on the unique strengths of the Swin Transformer, which uses a hierarchical

design and shifted window self-attention mechanism to capture both local and global information within aMRI images. This architecture, in conjunction with clinical data integration, provides a promising avenue for improving ASD classification accuracy compared to conventional deep learning models. By combining advanced neural network architectures with clinical data, this study seeks to push the boundaries of automated ASD diagnosis, ultimately contributing to more accurate and accessible tools for clinicians.

## 2. LITERATURE REVIEW

Alharthi, A, and Alzahrani, S.M. [1] explores the use of structural (sMRI) and functional MRI (fMRI) for diagnosing ASD, employing both 3D-CNN and vision transformers such as ConvNeXt, MobileNet, Swin, and ViT. The study involves generating 2D slices from 3D MRI scans across various brain planes (axial, coronal, and sagittal) and evaluates classification performance using the ABIDE dataset. Their findings indicate that multi-slice generation combined with advanced models leads to state-of-the-art accuracy, particularly when using 50-middle slices from fMRI data.

Honghao Cui et al. [2] describes a new approach combining CNN and transformer architectures to enhance glioma grading. The model, ResMT, integrates a spatial residual module (SRM) using 3D CNN and a Swin UNETR pre-trained segmentation model for improved tumor region analysis. The framework employs a multi-plane channel and spatial attention module (MCSA) for feature refinement across multiple planes. ResMT achieved an AUC of 0.9953 on the BraTs19 dataset, outperforming baseline models.

Zhentao Hu et al. [3] provide a solid reference for leveraging hybrid models in medical image analysis. This study emphasizes how the combination of convolutional neural networks (CNNs) and Transformers, especially with a shift window attention mechanism, improves feature extraction from MRI data. By addressing the challenge of local feature fusion in 3D medical images, Conv-Swinformer enhances the capture of fine-grained, lesion-specific features. It highlights the model's superior performance in capturing local details by focusing attention on small spatial areas, and layer-by-layer enlargement of the attention window improves semantic connection, offering better classification results for Alzheimer's disease. This paper could serve as a relevant for my own work on ASD diagnosis, where both CNNs and attention mechanisms might help in accurately identifying key patterns in MRI data.

Varun Ganjigunte Prakash et al. [4] proposed a novel method for human action recognition (HAR) in untrimmed videos, aimed at assessing autism spectrum disorder (ASD) in children. Their approach addressed the limitations of previous HAR models, such as performance degradation due to imprecise temporal region proposals and limited adaptability in clinical applications. The proposed behavior action recognition (BAR) pipeline incorporated child detection, temporal action localization, and identification of actions of interest. The model, trained on data from 400 children with ASD and 125 with other developmental delays (ODD), achieved diagnostic accuracies of 79.7% for ASD, 77.2% for ODD, and 80.8% for neurotypical children. Additionally, its performance on the Self-Stimulatory Behavior Dataset (SSBD) showed a top-1 accuracy of 78.57%, significantly higher than previous benchmarks for combined action recognition and localization.

Manu Gaur et al. [5] addressed key challenges in autism spectrum disorder (ASD) classification using deep learning models, focusing on scalability issues and poor out-of-distribution (OOD) performance in medical imaging scenarios. They highlighted that existing methods rely heavily on supervised learning, which demands large, annotated datasets that are costly to obtain, especially in neuroimaging. Furthermore, the performance of deep learning models often degrades in clinical settings due to the domain mismatch. To overcome these challenges, the authors proposed a self-supervised pretraining approach on in-domain data to enhance generalization and representation learning. They introduced an ensemble-based framework inspired by meta-embeddings in natural language processing, which integrates different self-supervised representations for vision tasks. By applying a 2D Discrete Fourier Transform, the framework captures global interactions among fused features, resulting in improved performance and robustness in medical imaging tasks with limited annotated data.

Fatima Zahra Benabdallah et al. [6] proposed a novel approach for enhancing the detection of autism spectrum disorder (ASD) using convolutional neural networks (CNN). The authors addressed challenges in achieving early and accurate ASD diagnosis by incorporating the theories of under- and over-connectivity deficits observed in the autistic brain. Their framework enhances connections related to these connectivity alterations in image-like connectivity matrices to aid in early diagnosis. Utilizing the multi-site Autism Brain Imaging Data Exchange (ABIDE I) dataset, the approach achieved a high prediction accuracy of up to 96%, indicating its potential for advancing ASD detection methods.

Deba Kanta Gogoi et al. [7] presented a deep learning approach to classify autism spectrum disorder (ASD) using MRI images. Their study focused on identifying structural and functional differences in the brain associated with ASD, which affect various aspects such as information processing and social cue interpretation. The research utilized a customized VGG16 architecture, along with other models like InceptionV3, ResNet50, DenseNet121, and MobileNet, to analyze brain MRI data from the publicly available ABIDE I dataset. After clustering the unlabeled data, the performance of the five deep learning models was compared. The findings demonstrated promising results in ASD classification, showing the potential of deep learning methods for analyzing brain MRI in the context of ASD.

Asrar G. Alharthi and Salha M. Alzahrani [8] conducted a comprehensive review of brain and vision transformers for autism spectrum disorder (ASD) diagnosis and classification. The paper discusses the challenges in identifying the causes and biomarkers of ASD, emphasizing the role of artificial intelligence in improving diagnostic capabilities. The authors review various MRI modalities and deep learning approaches, including convolutional neural networks (CNNs), autoencoders, graph convolutions, and attention networks, for ASD diagnosis. The review highlights the effectiveness of computer vision transformers, which often integrate CNN architectures and transfer learning techniques to enhance image classification performance. The paper also explores recent brain transformer

models, such as METAFormer, Com-BrainTF, Brain Network, and others, discussing their potential for ASD detection using transfer learning on MRI datasets. Additionally, the review suggests that specialized transformer-based models, inspired by natural language processing (NLP), could offer promising new directions for classifying ASD-related brain biomarkers by leveraging attention mechanisms and treating MRI data as sequence prediction tasks fine-tuned for brain disorder classification.

Xin Deng et al. [9] proposed a deep learning framework called spatial–temporal Transformer (ST-Transformer) to improve the classification of autism spectrum disorder (ASD) using time-series functional magnetic resonance imaging (fMRI) data. With the growing prevalence of ASD, there is a need for more objective and efficient diagnostic methods beyond traditional symptom-based clinical observation. The ST-Transformer employs a linear spatial–temporal multi-headed attention unit to capture the spatial and temporal patterns in fMRI data and incorporates a Gaussian GAN-based data balancing technique to address imbalances in real-world ASD datasets. The model was evaluated on two independent datasets (ABIDE I and ABIDE II), achieving accuracies of 71.0% and 70.6%, respectively, demonstrating competitive performance compared to existing state-of-the-art methods for ASD diagnosis.

Zhengning Wang et al. [10] introduced a self-attention deep learning framework for detecting autism spectrum disorder (ASD) and identifying structural biomarkers using morphological covariance brain networks. The study utilized structural MR images from the ABIDE consortium to classify ASD patients versus normal controls. The proposed framework leverages a transformer model to extract local and global features from individual structural covariance networks, enhancing the coordination patterns between brain regions compared to traditional CNN-based models. The self-attention coefficients map facilitated the identification of potential diagnostic biomarkers, primarily located in the prefrontal cortex, temporal cortex, and cerebellum. The method achieved a classification accuracy of 72.5% across various sites, outperforming many existing approaches. This research demonstrates the potential of self-attention deep learning frameworks in diagnosing ASD and establishing early biomarkers.

# 3. MATERIALS AND METHODS

## 3.1 DATA DESCRIPTION

The dataset for this study is sourced from the Autism Brain Imaging Data Exchange (ABIDE) repository, which contains neuroimaging data from individuals with autism spectrum disorder (ASD) and neurotypical controls. The ABIDE I dataset includes anatomical magnetic resonance imaging (aMRI) data, with T1-weighted structural MRI scans from 861 individuals diagnosed with ASD and 861 neurotypical controls across 17 international sites, covering a broad developmental age range from 7 to 35 years, thus enhancing the generalizability of the model. MRI scans were obtained using 3T MRI scanners with a voxel resolution of approximately 1 $mm^3$, producing high-resolution structural images. Preprocessing steps included skull-stripping, bias field correction for intensity non-uniformity, and spatial normalization to the MNI152 template for consistent orientation [11]). The 3D aMRI data were then converted into 2D slices for CNN input, selecting the middle slice along the z-axis to capture key anatomical structures, resized to 64x64 pixels to ensure uniformity. Both neuroimaging and clinical data, such as age, gender, and cognitive scores, were incorporated into the deep learning models to enhance classification accuracy. The clinical data file contains autism diagnostic scores, including ADI-R scores for social interaction, communication, and restricted behaviors, as well as ADOS scores for communication, social interaction, and stereotyped behaviors. Various models were used for classification, including a 3D CNN for spatial feature extraction, ResNet50, normal sequential CNN, 2D CNN with XGBoost, 2D CNN ResNet101, and Transformer-based models such as the Swin Transformer, aiming to improve ASD classification by integrating neuroimaging features with clinical data [12]).

## 3.2 MACHINE LEARNING AND DEEP LEARNING METHODS PROPOSED FOR ASD DETECTION

### *3.2.1 3D CNN -Resnet50Architecture for ASD Diagnosis:*

- ***Dataset Preparation***:

The dataset consists of T1-weighted structural MRI scans from the ABIDE repository, separated into ASD and non-ASD groups for both training and testing sets. Along with neuroimaging data, clinical data such as age, gender, and cognitive assessment scores were also included to improve model performance. The MRI scans were processed by extracting the middle slice of each 3D volume along the z-axis and resizing it to 224x224 pixels. These slices were converted from grayscale to RGB format to match the input requirements of ResNet50.

- ***Preprocessing***:

In addition to image preprocessing (resizing, grayscale-to-RGB conversion, and normalization), the clinical data, containing variables such as ADOS and ADI-R scores, was normalized and formatted for model integration. This clinical information includes scores related to social interaction, communication abilities, restricted and repetitive behaviors, and cognitive function, which are crucial diagnostic indicators for ASD. The data preprocessing pipeline was designed to shuffle and combine both the imaging and clinical data to create a balanced dataset for training and testing [13]).

- ***Model Architecture***:

A pre-trained ResNet50 model, designed to learn complex features from images, was employed for neuroimaging data. The model's weights were frozen to retain pre-learned features, and additional layers were added to fine-tune the classification task for ASD diagnosis. In parallel, the clinical data (including age, gender, ADOS, and ADI-R scores) was integrated into the model by concatenating it with the features extracted from the MRI images. This approach allowed the model to consider both anatomical features from the brain and important clinical factors when making its predictions. The classification head consisted of GlobalAveragePooling2D and Dense layers to combine both data types, concluding with a sigmoid-activated output layer for binary classification (ASD or non-ASD).

• ***Training Process***:

The combined model, incorporating both neuroimaging and clinical data, was trained using binary cross-entropy as the loss function, the Adam optimizer, and accuracy as the primary metric. The model was trained for 10 epochs on the dataset, with a split for training and validation. By including clinical data in the training process, the model aimed to improve its ability to distinguish between ASD and non-ASD subjects.

• ***Evaluation***:

The model's performance was evaluated on a separate test set consisting of MRI slices and corresponding clinical data. This evaluation demonstrated the model's ability to generalize to unseen data, highlighting the role of clinical variables in enhancing the classification accuracy. The test set accuracy and loss were recorded, providing a measure of the model's effectiveness in ASD detection.

• ***Prediction on New Data***:

The model was further evaluated using unseen MRI images and clinical data for individual predictions. A new MRI image was loaded and preprocessed (including the same middle slice extraction and resizing). The corresponding clinical data, including age and ADOS/ADI-R scores, was also input into the model. The combined neuroimaging and clinical data provided a robust input for the model, resulting in a final prediction based on both anatomical and clinical information. The prediction value was recorded, offering insight into the model's confidence in its ASD or non-ASD classification [14].

#### *3.2.2 3D Normal CNN -Sequential Architecture for ASD Diagnosis:*

• ***Dataset Preparation***:

The dataset comprises 3D structural MRI scans along with clinical data, including variables such as age, gender, and cognitive assessment scores, sourced from relevant directories. The MRI scans were separated into ASD and non-ASD groups for both training and testing. The images were preprocessed by extracting and resizing each 3D volume to a uniform size of 64x64x64 pixels. Additionally, the clinical data was formatted to ensure compatibility with the neuroimaging data, facilitating a comprehensive analysis that integrates both data types for improved model performance.

• ***Preprocessing***:

Preprocessing involved several key steps for both MRI images and clinical data. The MRI images were loaded, resized, and normalized to ensure pixel intensity values ranged between 0 and 1. Each image was expanded to include a channel dimension, preparing them for input into the 3D CNN model. For the clinical data, normalization was performed on the age, gender, and cognitive scores to maintain a consistent scale. This clinical information, which includes essential indicators for ASD diagnosis, was then combined with the imaging data to create a balanced dataset that reflects both anatomical features and relevant clinical factors [15].

• ***Model Architecture***:

The model architecture utilized a 3D Convolutional Neural Network (CNN) designed to learn spatial features from the MRI volumes. A Sequential model was built using the following layers:

- Conv3D Layer: The first layer consists of 32 filters with a kernel size of (3, 3, 3) and uses ReLU activation to capture local features.
- MaxPooling3D Layer: A pooling layer with a pool size of (2, 2, 2) follows to reduce dimensionality and retain the most significant features.
- Conv3D Layer: The second convolutional layer has 64 filters, again with a kernel size of (3, 3, 3) and ReLU activation.
- MaxPooling3D Layer: Another pooling layer to further downsample the feature maps.
- Conv3D Layer: A third convolutional layer with 128 filters and ReLU activation.
- MaxPooling3D Layer: A pooling layer to reduce dimensions.
- GlobalAveragePooling3D Layer: This layer averages the spatial dimensions, reducing the feature maps to a fixed size.
- Dense Layer: A fully connected layer with 128 units and ReLU activation for additional learning.
- Dropout Layer: A dropout layer with a rate of 0.5 to prevent overfitting.
- Dense Output Layer: The final layer has a single unit with a sigmoid activation function for binary classification (ASD vs. non-ASD).

The model integrates clinical data by concatenating it with the features extracted from the MRI scans, allowing for a comprehensive analysis that combines both neuroimaging and clinical factors.

• ***Training Process***:

The combined model, incorporating both neuroimaging and clinical data, was trained using binary cross-entropy as the loss function, the Adam optimizer, and accuracy as the primary evaluation metric. The training was conducted for 10 epochs, using a dataset split for training and validation. By including clinical data in the training process, the model aimed to enhance its ability to distinguish between ASD and non-ASD subjects, leveraging the complementary information provided by the clinical variables.

• ***Evaluation***:

The model's performance was evaluated on a separate test set that included both MRI volumes and corresponding clinical data. The evaluation focused on the model's ability to generalize to unseen data, with metrics such as test accuracy and loss recorded. This assessment demonstrated the model's effectiveness in classifying ASD and non-ASD subjects and highlighted the role of clinical variables in enhancing classification performance, suggesting that the inclusion of clinical data can lead to more accurate predictions [16].

• ***Prediction on New Data***:

For individual predictions, a new MRI volume and its corresponding clinical data (e.g., age and cognitive scores) were loaded and preprocessed using the same methods as the training data. The model predicted the class of the new data based on the combined input from both neuroimaging and clinical information. The prediction value was recorded, indicating the model's confidence in classifying the input as either ASD or non-ASD.

This integrated approach underscores the importance of considering both anatomical and clinical factors when making diagnostic predictions.

#### *3.2.3 2D CNN-Combined XgBoost Architecture for ASD Diagnosis:*

- ***Dataset Preparation***:

The dataset used consists of structural MRI scans from two groups: ASD and non-ASD subjects, for both training and testing purposes. These MRI scans were sourced and divided into respective classes. Alongside the neuroimaging data, clinical data including key diagnostic scores such as ADOS and ADI-R was incorporated to improve classification accuracy. The neuroimaging data was preprocessed by extracting the central slice from each 3D MRI volume along the z-axis, resized to 224x224 pixels, and converted from grayscale to RGB to match the input requirements of the ResNet50 model.

- ***Preprocessing***:

The preprocessing pipeline involved multiple steps for both neuroimaging and clinical data. For the MRI images, after resizing and conversion to RGB, the pixel values were normalized by scaling them to a range of [0, 1]. In parallel, the clinical data was loaded and processed, focusing on variables such as ADI-R social and communication scores, ADOS totals, and other diagnostic indicators for ASD. These clinical features were normalized to ensure consistency with the image data. Both datasets were shuffled and split for training and testing, creating a balanced dataset for model training and evaluation [17].

- ***Clinical Data***:

The clinical data file contains crucial information related to autism spectrum disorder (ASD) assessments, providing valuable diagnostic scores for each subject. This data includes the SUB_ID, which serves as a unique identifier for each subject, and several key measures from the Autism Diagnostic Interview-Revised (ADI-R), such as ADI_R_SOCIAL_TOTAL_A for social interaction difficulties, ADI_R_VERBAL_TOTAL_BV for verbal communication issues, and ADI_R_NONVERBAL_TOTAL_BV for non-verbal communication challenges. Other ADI-R scores include ADI_R_RRB_TOTAL_C for restricted and repetitive behaviors, ADI_R_ONSET_TOTAL_D for age of onset of symptoms, and ADI_R_RSRCH_RELIABLE, which indicates the research reliability of the ADI-R data. Additionally, the Autism Diagnostic Observation Schedule (ADOS) provides several important columns, including ADOS_MODULE for the specific module used, ADOS_RSRCH_RELIABLE for ADOS data reliability, and scores like ADOS_G_TOTAL (overall score), ADOS_G_COMM (communication difficulties), ADOS_G_SOCIAL (social interaction challenges), and ADOS_G_STEREO_BEHAV (stereotyped behaviors). This detailed clinical data enhances the model's ability to make accurate ASD classifications [18].

- ***Model Architecture***:

A combined model architecture was employed, using a pre-trained ResNet50 model to handle the MRI data. The ResNet50's convolutional layers, pre-trained on ImageNet, were frozen to retain the learned visual features. Simultaneously, the clinical data was processed using a separate fully connected (dense) layer. The outputs from both the ResNet50 feature extraction and the clinical data pipeline were then concatenated, allowing the model to learn from both neuroanatomical and clinical information. The combined representation was passed through additional dense layers, followed by a final sigmoid-activated output for binary classification (ASD or non-ASD).

- ***Training Process***:

The combined model, integrating both neuroimaging and clinical data, was trained using the Adam optimizer and binary cross-entropy as the loss function. The model was trained for a fixed number of epochs (20) with early stopping to prevent overfitting. Data augmentation techniques, such as rotations, zooms, and flips, were applied to the MRI images during training to improve model robustness. Cross-validation using a 5-fold approach was utilized to ensure the model generalizes well across different subsets of the data.

- ***Evaluation***:

The model was evaluated on a separate test set, containing MRI slices and corresponding clinical data. The combined model's performance was measured in terms of test accuracy and loss, demonstrating its ability to generalize well to unseen data. Additionally, the performance of an XGBoost classifier trained solely on the clinical data was evaluated for comparison [19]. The accuracy results from both models were recorded, and the combined predictions from ResNet50 and XGBoost were averaged for final classification.

- ***Prediction on New Data***:

The trained model was used to make predictions on new MRI scans and clinical data, providing final classifications. MRI images were preprocessed similarly by extracting and resizing the middle slice, while clinical data was formatted to match the input structure. The model then combined both neuroimaging and clinical inputs to make predictions, offering insight into whether a given subject falls under the ASD or non-ASD category based on the model's output. The final predictions provided a comprehensive diagnosis by leveraging both structural brain data and clinical assessments.

#### *3.2.4 2D CNN Resnet101 Architecture for ASD Diagnosis:*

- ***Dataset Preparation***:

The dataset comprises T1-weighted structural MRI scans categorized into ASD and non-ASD groups, used for both training and testing. Alongside neuroimaging data, clinical information such as diagnostic scores from assessments like ADOS and ADI-R were included. MRI scans were processed by extracting the middle slice from each 3D volume and resizing it to 224x224 pixels to match ResNet101's input size. These slices were used for the model, representing anatomical features critical for ASD classification.

- ***Preprocessing***:

The neuroimaging data underwent preprocessing steps, including resizing and normalization. The clinical data, containing diagnostic scores such as ADOS (communication, social interaction, and stereotyped behaviors) and ADI-R (social, verbal, non-verbal difficulties, and repetitive behaviors), was normalized to ensure consistency with the model's input. Both imaging and clinical data were shuffled and combined to create a balanced training and testing dataset [20].

• ***Model Architecture***:

The model used a pre-trained ResNet101 architecture to extract complex features from the MRI data. The ResNet101 weights were frozen to retain their learned image-based features, and custom layers were added to fine-tune the model for ASD diagnosis. Simultaneously, clinical data, such as ADOS and ADI-R scores, was processed and concatenated with features extracted from MRI slices. This combined approach allowed the model to consider both neuroanatomical and clinical characteristics when predicting ASD. The final classification layer included a sigmoid activation function for binary classification (ASD or non-ASD).

• ***Training Process***:

The model was trained using a combination of neuroimaging and clinical data, with binary cross-entropy as the loss function and the Adam optimizer. Training was performed over 10 epochs with a training-validation split. The inclusion of clinical data aimed to boost the model's ability to differentiate between ASD and non-ASD subjects more effectively. K-fold cross-validation was used to ensure robust performance across multiple data splits, and data augmentation techniques such as rotation and zoom were applied to enhance generalization.

• ***Evaluation***:

The trained model was evaluated on a separate test dataset comprising MRI slices and corresponding clinical features. The evaluation measured the model's ability to generalize to unseen data and highlighted the positive impact of incorporating clinical data in enhancing prediction accuracy. Test set accuracy, loss, and other performance metrics were recorded to assess model effectiveness in diagnosing ASD [21].

• ***Prediction on New Data***:

The model was further tested using unseen MRI images along with associated clinical data for individual predictions. A new MRI scan was processed by extracting and resizing its middle slice, and the corresponding clinical features, including diagnostic scores, were fed into the model. The combined input (MRI and clinical) allowed the model to generate a prediction, providing insight into the likelihood of the subject having ASD based on both anatomical and clinical indicators. The prediction confidence score was recorded to gauge the model's performance on new cases.

### *3.2.5 Vision Transformers (ViT) Architecture for ASD Diagnosis:*

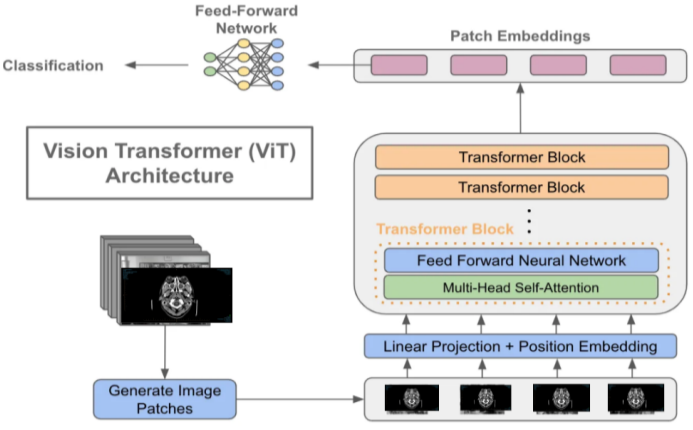


Fig.1. ViT Architecture diagram

The Vision Transformer (ViT) represents a significant shift in how deep learning models are applied to image processing and computer vision tasks. Originally developed for Natural Language Processing (NLP), transformer architectures have been adapted for visual data, demonstrating remarkable effectiveness in a variety of image recognition tasks.

- **Architecture Overview:** The key innovation of the ViT lies in its ability to treat images as sequences of patches, similar to how transformers handle sequences of words. Instead of relying solely on convolutional layers, ViT divides an input image into smaller fixed-size patches, which are then linearly embedded into a sequence of tokens (Fig.1). This sequence is fed into a standard transformer architecture that employs self-attention mechanisms to capture complex relationships between different parts of the image [22].
- **Self-Attention Mechanism:** The self-attention mechanism allows the model to weigh the importance of each patch relative to others, enabling it to learn contextual information and dependencies effectively. This capability is crucial for understanding spatial relationships and details within an image, making ViT particularly powerful for tasks that require a holistic understanding of visual content.
- **Training and Performance:** ViT has been pre-trained on large datasets, such as ImageNet, leveraging transfer learning to adapt its learned features to specific tasks. By fine-tuning the model on a smaller dataset tailored to a specific application (like ASD classification in neuroimaging), ViT can achieve state-of-the-art performance, often outperforming traditional convolutional neural networks (CNNs) in various visual recognition challenges.
- **Advantages and Applications:** One of the primary advantages of ViT is its scalability; as the model size increases, its performance improves significantly, making it suitable for a range of applications from medical imaging to autonomous driving. In the context of neuroimaging, ViT can analyze structural MRI scans, providing valuable insights into brain anatomy and associated disorders, such as ASD [23].
- ***Building the Combined Model with Vision Transformer (ViT)***:

The combined model utilizes a pre-trained Vision Transformer (ViT) to analyze MRI scan data, enhancing the classification of subjects as either ASD or non-ASD. ViT, an advanced deep learning architecture, applies transformer mechanisms, initially designed for Natural Language Processing (NLP), to visual data, achieving state-of-the-art performance in various image recognition tasks. The model is initialized by loading the ViT architecture, which is pre-trained on a large dataset. Importantly, the model excludes the top classification layers to allow for customization, focusing specifically on binary classification [24].

To retain the learned features from pre-training, the ViT layers are frozen during the training process, ensuring that the valuable representations acquired from the initial training phase are preserved. This freezing step is crucial for leveraging the rich feature extraction capabilities of the ViT without introducing unnecessary variability from fine-tuning.

• ***Image Input Pipeline***:

MRI scans are processed as inputs to the model, resized to the appropriate dimensions (224x224 pixels) to meet the requirements of the ViT architecture. The images are input through a dedicated pipeline where they are fed into the pre-trained ViT model. Following this, a Global Average Pooling layer is applied to condense the feature maps generated by the ViT into a smaller, manageable feature vector. This pooling step is essential for facilitating the subsequent integration of clinical data into the model.

• ***Clinical Data Input Pipeline***:

Alongside the MRI data, clinical information relevant to ASD diagnosis—such as scores from standardized assessments like the ADOS and ADI-R—is processed through a separate input pipeline. This clinical data is fed into the model via a simple feedforward neural network. The clinical features undergo transformation through a dense layer, utilizing ReLU activation to create a feature vector that can seamlessly integrate with the image-derived features from the ViT.

• ***Combining Image and Clinical Features***:

The outputs from the image processing pipeline and the clinical data pipeline are concatenated to form a unified input for the final stages of the model. This concatenation allows the model to simultaneously consider the anatomical information derived from the MRI images and the pertinent clinical data when making predictions about ASD classification. The combined feature representation is then passed through additional dense layers to refine the model's predictions, culminating in a final output layer that employs a sigmoid activation function for binary classification [25].

• ***Model Compilation and Training***:

The combined model is compiled using the Adam optimizer and binary cross-entropy as the loss function, standard choices for addressing binary classification tasks. The training process incorporates both the neuroimaging and clinical data, with early stopping implemented to halt training when performance on the validation set no longer improves. This strategy prevents overfitting and ensures that the model generalizes well to new data.

The model undergoes training over a specified number of epochs, with the option for early stopping to enhance the efficiency and effectiveness of the learning process. After training, the model is saved for future predictions.

• ***Loading and Preprocessing New MRI Data for Prediction***:

Once trained, the model can be reloaded to make predictions on new, unseen MRI data. This involves preprocessing the MRI scans using libraries designed for neuroimaging formats, such as nibabel. The preprocessing includes extracting the middle slice of each 3D MRI volume and resizing it to fit the model's input specifications. This consistent preprocessing ensures that new data aligns with the training data format, enabling accurate predictions.

• ***Predicting ASD and Non-ASD***:

After preprocessing the MRI images and generating corresponding clinical data (often using dummy values for testing), the model makes predictions regarding whether a subject falls into the ASD or non-ASD category. The model combines both neuroimaging and clinical inputs to produce a prediction score. Based on this score, thresholds are applied (e.g., values above 0.5 indicate ASD, while those below indicate non-ASD) to determine the final classification.

Finally, the model tracks and outputs the total counts of ASD and non-ASD predictions across the set of test images. This summary provides insight into the model's performance and efficacy in classifying subjects based on both neuroimaging and clinical data. The integrated approach of utilizing Vision Transformers alongside clinical assessments establishes a robust framework for ASD classification, with potential for further refinement and enhancement as more data becomes available.

### *3.2.6 Swin Transformer Architecture for ASD Diagnosis*

The Swin Transformer is a cutting-edge deep learning architecture designed for visual tasks, extending the principles of the original Transformer model to image processing. Unlike traditional vision models, it employs a hierarchical approach that processes images in a sequence of non-overlapping local windows, allowing for efficient modeling of both local and global contextual information. The architecture divides the input image into patches and computes self-attention within these windows, progressively merging neighboring patches at different stages to form a pyramid-like structure [26]. This hierarchical representation enables Swin Transformer to capture fine-grained details and high-level semantic information, making it well-suited for complex classification tasks like ASD diagnosis using neuroimaging data (Fig.2).

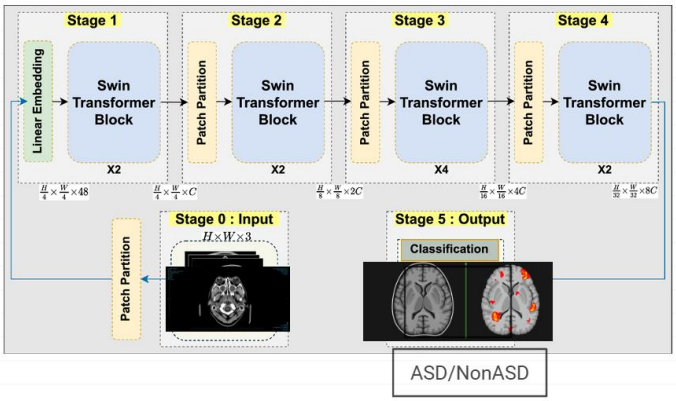


Fig.2. Swin Transformer Architecture

• ***Building the Combined Model with Swin Transformer***:

The combined model utilizes a Swin Transformer architecture to analyze MRI scan data for ASD classification. Swin Transformer is an advanced vision transformer-based model that uses a hierarchical approach to extract local and global features. The model was initialized by defining a Swin Transformer architecture with additional layers for image and clinical data processing. The Swin Transformer acts as the core feature extractor from the MRI images, while clinical data is integrated using a separate neural network branch.

• ***Image Input Pipeline***:

MRI scans were preprocessed and resized to dimensions of 224x224 pixels to fit the Swin Transformer's input specifications. Each 3D scan was converted into a 2D slice by selecting the middle slice along the z-axis, capturing key anatomical structures.

The preprocessing involved normalizing the pixel values to the range [0, 1]. The image input pipeline fed the preprocessed MRI scans into the Swin Transformer, where a Global Average Pooling layer condensed the feature maps for integration with clinical data [27].

• ***Clinical Data Input Pipeline***:

Clinical data, including various scores from standardized ASD assessments (ADI-R, ADOS), were processed through a separate input pipeline. This branch used a feedforward neural network to transform the clinical features into a feature vector using dense layers with ReLU activation. This transformation helped to capture relevant information from the clinical data, which was then integrated with the MRI-derived features [28].

• ***Combining Image and Clinical Features***:

The model concatenated the outputs from the image processing (Swin Transformer) and clinical data pipelines to form a unified representation. This combined feature vector was passed through additional dense layers, which performed further refinement before the final output layer. The final classification layer used a sigmoid activation function to predict the probability of ASD, allowing for binary classification (ASD vs. non-ASD).

• ***Model Compilation and Training***:

The model was compiled using the Adam optimizer and binary cross-entropy as the loss function, suitable for binary classification. A K-Fold Cross-Validation approach was adopted to evaluate model performance across different data splits, helping to ensure robust generalization. Early stopping was used during training to prevent overfitting, with the model saved in TensorFlow's SavedModel format for future predictions.

• ***Loading and Preprocessing New MRI Data for Prediction***:

For predictions on new MRI data, the model expected inputs to follow the same preprocessing steps applied during training. The MRI images were processed using nibabel to extract the middle slice and resized to 224x224 pixels. This consistency in preprocessing ensured that the model would generalize well to unseen data.

• ***Predicting ASD and Non-ASD***:

The trained model was used to predict the classification of new subjects by processing both MRI data and associated clinical features. Predictions were based on combining anatomical information derived from the MRI scans and clinical scores, producing a probability score indicating whether the subject had ASD [29].

#### *3.2.7 Simplified CNN-Swin Transformer Hybrid Model for ASD Diagnosis:*

This model enhances the Simplified CNN-Swin Transformer Hybrid Model by incorporating clinical data alongside MRI features. By combining imaging data with relevant clinical information, such as ASD-related assessment scores, the model improves its ability to classify ASD with greater accuracy [30]. The model's architecture was modified to include a dual input system where MRI data is processed through a CNN-based pipeline, and clinical data is processed through a separate neural network branch:

1. **Image Feature Extraction (CNN)**:
- MRI data is processed through convolutional layers to extract spatial features.
- Global average pooling condenses the feature maps, yielding a compact image feature representation.

2. **Clinical Data Processing**:
- Clinical data, including standardized ASD assessment scores (e.g., ADI-R, ADOS), is passed through a feedforward neural network.
- This branch uses dense layers with ReLU activation to capture clinical patterns related to ASD.
- The clinical data is transformed into a feature vector, which complements the MRI-derived features.

3. **Feature Concatenation and Final Classification**:
- The outputs from the CNN and clinical data branches are concatenated to form a unified feature vector.
- This combined representation is passed through additional dense layers, which perform further feature refinement.
- The final classification layer with sigmoid activation predicts the probability of ASD, enabling binary classification (ASD vs. non-ASD).

#### *3.2.8 Data Preparation and Preprocessing:*

**MRI Data Pipeline**

- **2D Slice Extraction**: A middle slice of the MRI image is selected along the z-axis for input.
- **Normalization**: Pixel values are normalized to [0, 1].
- **Image Augmentation**: Standard transformations (rotation, shift, zoom, etc.) are applied to improve model robustness.

**Clinical Data Pipeline**

- **Clinical Scores**: Various ASD-related scores are preprocessed and normalized, making them compatible with the model's input.
- **Feature Scaling**: Clinical features are scaled to ensure they are on a comparable range to the MRI features.

**3. Model Training and Validation**

The training setup involved:

- **Adam Optimizer**: Used for stable training with binary cross-entropy as the loss function.
- **K-Fold Cross-Validation**: Validated the model across multiple data splits to improve generalization.
- **Early Stopping and Learning Rate Scheduling**: These callbacks were applied to prevent overfitting and to dynamically adjust the learning rate during training.

**4. Model Evaluation**

After training, the model was evaluated on a separate test set:

- **Combined Feature Evaluation**: MRI and clinical data features were used jointly, allowing for more context-aware classification of ASD.
- **Improved Accuracy**: By integrating both data types, the model demonstrated a significant improvement in predictive accuracy over models trained on MRI data alone.

The model was saved in TensorFlow's SavedModel format to facilitate future predictions. Consistent preprocessing of MRI and

clinical data ensures the model performs well on new, unseen data.

## 4. EXPERIMENTAL RESULTS

Deep learning models were trained on the input data for children's behavior classification in order to extract the abnormal behavior traits that indicate autism. The models were trained on an AMD 7th Generation HP Victus Ryzen laptop with NVidia GeForce GTX 3050 GPU. We utilized the Python programming language in conjunction with prominent deep learning libraries, namely TensorFlow and Keras.

### 4.1 EVALUATION METRICS

A classification report is a statistical measurement of performance in the deep learning field. Its objective is to demonstrate the performance of the training classification model, including its accuracy, recall, F1 score, and overall support [31].

- *Accuracy* is defined as the ratio of correct predictions to the total number of predictions made.
- *Precision* is defined as the proportion of true positives to the total number of true and false positives in a particular sample.
- *Recall* is defined as the proportion of true positives to the sum of true positives and false negatives.
- The *F1 score* is a weighted harmonic mean of accuracy and recall [31]. When the F1 score is closer to the number 1.0, it means the model has high performance.

Table.1. Results

| Method (with Clinical Data) | Accuracy (%) |
|---|---|
| 3D CNN ResNet50 | 62 |
| Normal CNN (Sequential) | 69 |
| 2D CNN + XGBoost | 78 |
| 2D CNN ResNet101 | 60 |
| Transformer | 70 |
| Swin Transformer | 75 |
| Hybrid CNN + Swin | 80 |

### 4.2 PROPOSED HYBRID CNN AND SWIN TRANSFORMER ADVANTAGES

The proposed hybrid model, which integrates a simplified CNN with features of a Swin Transformer, offers significant advantages in the analysis of clinical data, particularly in medical image classification tasks. CNNs excel in extracting spatial hierarchies of features from images, effectively capturing local patterns crucial for analyzing 2D slices of medical imaging data, such as NIfTI files. This capability is essential for identifying subtle variations in medical conditions, such as ASD versus non-ASD cases, based on MRI scans. In addition to this, the model incorporates a Swin Transformer, renowned for its ability to capture long-range dependencies and contextual information, which enhances its capacity to understand the overall structure and relationships within the images.

This integration allows the model to focus on critical areas while maintaining a broader view of the entire image, essential for clinical applications where detailed differentiation between conditions can impact diagnosis and treatment. Furthermore, the incorporation of clinical metadata alongside image data, such as patient demographics, clinical history, and behavioral assessments, can improve model performance by providing contextual background that aids in making more informed predictions [32].

By using mixed precision training, the model optimizes memory usage, facilitating the handling of larger datasets common in clinical settings and enabling efficient computation. Data augmentation techniques employed during training further enhance generalization, ensuring the model performs well across varied patient data and imaging conditions. The result is a robust model that effectively combines the local feature extraction strength of CNNs with the global context understanding provided by Transformers, thereby creating a powerful tool for medical image analysis that can lead to more accurate and reliable clinical outcomes.

Table.2. Confusion Matrix of the proposed Model

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Non-ASD | 0.83 | 0.76 | 0.80 | 46 |
| ASD | 0.78 | 0.85 | 0.81 | 46 |
| Accuracy | | | 0.80 | 92 |
| Macro Avg. | 0.81 | 0.80 | 0.80 | 92 |
| Weighted Avg. | 0.81 | 0.80 | 0.80 | 92 |

The Table.2 shows the performance metrics of a classification model for two classes: Non-ASD and ASD. Here's what each metric represents:

### 4.3 CLASS-SPECIFIC METRICS

1. **Precision**: Measures the accuracy of positive predictions for each class.

- For **Non-ASD**: 0.83, meaning 83% of the instances predicted as Non-ASD are actually Non-ASD.
- For **ASD**: 0.78, meaning 78% of the instances predicted as ASD are actually ASD.

2. **Recall**: Measures the model's ability to correctly identify all actual positives.

- For **Non-ASD**: 0.76, indicating that the model correctly identified 76% of the true Non-ASD cases.
- For **ASD**: 0.85, indicating that the model correctly identified 85% of the true ASD cases.

3. **F1-Score**: The harmonic mean of precision and recall, which provides a balance between the two.

- For **Non-ASD**: 0.80, suggesting good overall accuracy for this class.
- For **ASD**: 0.81, showing similar performance to Non-ASD.

4. **Support**: The number of instances in each class.

- Both **Non-ASD** and **ASD** have 46 instances, indicating a balanced dataset.

#### *4.3.1 Overall Metrics:*

- **Accuracy**: The proportion of correctly classified instances across all samples. **0.80** (80%), meaning the model correctly classified 80% of the total samples.
- **Macro Avg**: The unweighted average of precision, recall, and F1-score across all classes. Precision, recall, and F1-score are all 0.80–0.81, reflecting balanced performance across both classes.
- **Weighted Avg**: The average of precision, recall, and F1-score, weighted by the number of instances in each class. Similar to macro average (0.80–0.81), indicating the model's consistent performance even when considering the class distribution.

Overall, the model shows good performance in detecting both ASD and Non-ASD cases, with a slight bias towards identifying ASD more accurately (higher recall for ASD).

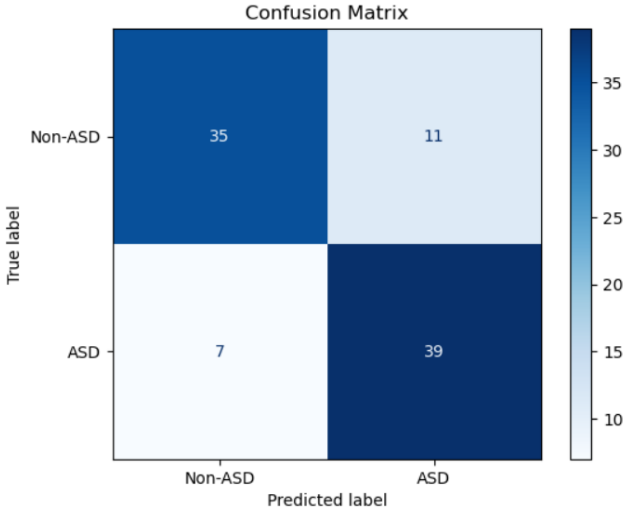


Fig.3. Confusion Matrix

### 4.4 HYBRID CNN+SWIN TRANSFORMER

The confusion matrix illustrates the performance of a classification model on the ASD and Non-ASD classes. Here's a breakdown:

- **True Positives (ASD correctly classified)**: 39 instances where the model predicted ASD, and the true label was also ASD.
- **True Negatives (Non-ASD correctly classified)**: 35 instances where the model predicted Non-ASD, and the true label was Non-ASD.
- **False Positives (Non-ASD incorrectly classified as ASD)**: 11 instances where the model predicted ASD, but the true label was Non-ASD.
- **False Negatives (ASD incorrectly classified as Non-ASD)**: 7 instances where the model predicted Non-ASD, but the true label was ASD.
- The model achieved good classification for both classes, with higher true positives (39) and true negatives (35), indicating a relatively balanced performance.
- There are some errors, with more false positives (11) than false negatives (7), suggesting a slight bias towards predicting ASD when it is actually Non-ASD. The overall accuracy, as calculated from the matrix, is 80%

## 5. CONCLUSION AND FUTURE WORK

In this study, we evaluated the effectiveness of integrating neuroimaging data and clinical assessments for diagnosing ASD using deep learning models. Our results demonstrate that hybrid models, especially the Hybrid CNN+Swin Transformer, achieve the highest classification accuracy at 80%, outperforming traditional CNN architectures and pure transformer-based approaches. This suggests that combining the spatial feature extraction strengths of CNNs with the Swin Transformer's global contextual understanding offers a more comprehensive approach for identifying ASD-related patterns in brain imaging data. Despite these promising results, there are limitations to address in future research. The relatively small sample size and specific imaging data used in this study may limit the generalizability of the findings.

Future efforts should focus on expanding the dataset with more diverse samples and incorporating multimodal neuroimaging data, such as fMRI, to improve the model's robustness. Additionally, integrating other data types, including clinical and genetic information, may further enhance predictive accuracy and contribute to a more personalized diagnostic approach.

Future studies should also emphasize improving model interpretability to better understand the neuroanatomical patterns linked to ASD. Utilizing attention mechanisms within transformers could help visualize brain regions that are critical for model predictions, providing transparency that could aid clinicians in understanding and validating the model's decisions. This advancement would help pave the way for more reliable AI-assisted diagnostics in clinical settings. Finally, ongoing research efforts are exploring the potential of integrating Generative AI with CNN and Swin Transformer models, which may open new avenues for enhancing ASD classification.

## REFERENCES

[1] A.G. Alharthi and S.M. Alzahrani, "Multi-Slice Generation sMRI and fMRI for Autism Spectrum Disorder Diagnosis Using 3D-CNN and Vision Transformers", *Brain Sciences*, Vol. 13, No. 11, pp. 1578-1589, 2023.

[2] H. Cui, Z. Ruan, Z. Xu, X. Luo, J. Dai and D. Geng, "ResMT: A Hybrid CNN-Transformer Framework for Glioma Grading with 3D MRI", *Computers and Electrical Engineering*, Vol. 120, pp. 1-6, 2024.

[3] Z. Hu, Y. Li, Z. Wang, S. Zhang, W. Hou and Alzheimer's Disease Neuroimaging Initiative, "Conv-Swinformer: Integration of CNN and Shift Window Attention for Alzheimer's Disease Classification", *Computers in Biology and Medicine*, Vol. 164, pp. 1-6, 2023.

[4] V.G. Prakash, M. Kohli, A.P. Prathosh, M. Juneja, M. Gupta, S. Sairam and N. Goyal, "Video-based Real-Time Assessment and Diagnosis of Autism Spectrum Disorder using Deep Neural Networks", *Expert Systems*, pp. 1-6, 2023.

[5] M. Gaur, K. Chaturvedi, D.K. Vishwakarma, S. Ramasamy and M. Prasad, "Self-Supervised Ensembled Learning for Autism Spectrum Classification", *Research in Autism Spectrum Disorders*, Vol. 107, pp. 1-9, 2023.

[6] F.Z. Benabdallah, A. Drissi El Maliani, D. Lotfi and M. El Hassouni, "A Convolutional Neural Network-Based Connectivity Enhancement Approach for Autism Spectrum Disorder Detection", *Journal of Imaging*, Vol. 9, No. 6, pp. 1-9, 2023.

[7] D.K. Gogoi, J. Talukdar, D.K. Bhattacharyya and T.P. Singh, "A Deep Learning Approach to Classify Autism Spectrum Disorder using MRI Images", Vol.12, pp. 1-7, 2023.

[8] W. Jiang, S. Liu, H. Zhang, X. Sun, S.H. Wang, J. Zhao and J. Yan, "CNNG: A Convolutional Neural Network with Gated Recurrent Units for Autism Spectrum Disorder Classification", *Frontiers in Aging Neuroscience*, Vol. 14, pp. 1-8, 2022.

[9] X. Deng, J. Zhang, R. Liu and K. Liu, "Classifying ASD based on Time-Series fMRI using Spatial-Temporal Transformer", *Computers in Biology and Medicine*, Vol. 151, pp. 1-9, 2022.

[10] Z. Wang, D. Peng, Y. Shang and J. Gao, "Autistic Spectrum Disorder Detection and Structural Biomarker Identification using Self-Attention Model and Individual-Level Morphological Covariance Brain Networks", *Frontiers in Neuroscience*, Vol. 15, pp. 1-6, 2021.

[11] A. Di Martino, C.G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts and M.P. Milham, "The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism", *Molecular Psychiatry*, Vol. 19, No. 6, pp. 659-667, 2014.

[12] J.M. Lee, S. Kyeong, E. Kim and K.A. Cheon, "Abnormalities of Inter-and Intra-Hemispheric Functional Connectivity in Autism Spectrum Disorders: A Study using the Autism Brain Imaging Data Exchange Database", *Frontiers in Neuroscience*, Vol. 10, pp.1-7, 2016.

[13] J. Deng, M.R. Hasan, M. Mahmud, M.M. Hasan, K.A. Ahmed and M.Z. Hossain, "Diagnosing Autism Spectrum Disorder using Ensemble 3D-CNN: A preliminary Study", *Proceedings of International Conference on Image Processing*, pp. 3480-3484, 2022.

[14] M.R. Lamani and P. Julian Benadit, "A Review on Deep Learning Algorithms in the Detection of Autism Spectrum Disorder", *Congress on Intelligent Systems*, pp. 283-297, 2023.

[15] S. Mostafa and F.X. Wu, "Diagnosis of Autism Spectrum Disorder with Convolutional Autoencoder and Structural MRI Images", *Neural Engineering Techniques for Autism Spectrum Disorder*, pp. 23-38, 2021.

[16] G. Li, M. Liu, Q. Sun, D. Shen and L. Wang, "Early Diagnosis of Autism Disease by Multi-Channel CNNs", *Proceedings of International Conference on Machine Learning in Medical Imaging*, pp. 303-309, 2018.

[17] S. Akter, H. Shahriar and A. Cuzzocrea, "Autism Disease Detection using Transfer Learning Techniques: Performance Comparison between Central Processing Unit vs Graphics Processing Unit Functions for Neural Networks", *Proceedings of International Conference on Computers, Software and Applications*, pp. 1084-1092, 2023.

[18] J.B. Lebersfeld, M. Swanson, C.D. Clesi and S.E. O'Kelley, "Systematic Review and Meta-Analysis of the Clinical Utility of the ADOS-2 and the ADI-R in Diagnosing Autism Spectrum Disorders in Children", *Journal of Autism and Developmental Disorders*, pp. 1-14, 2021.

[19] K. Mittal, K.S. Gill, K. Rajput and V. Singh, "Utilizing Machine Learning and Employing the XGBoost Classification Technique for Evaluating the Likelihood of Autism Spectrum Disorder", *Proceedings of International Conference on Emerging Technology*, pp. 1-5, 2024.

[20] A. Toranjsimin, S. Zahedirad and M.H. Moattar, "Robust Low Complexity Framework for Early Diagnosis of Autism Spectrum Disorder based on Cross Wavelet Transform and Deep Transfer Learning", *SN Computer Science*, Vol. 5, No. 2, pp. 1-6, 2024.

[21] M.M. Rashid, M.S. Alam, M.A. Haque, M.Y. Ali and S. Yvette, "Effect of Different Modalities of Facial Images for Diagnosis of ASD by Deep Neural Network", *AIP Conference Proceedings*, Vol. 3161, No. 1, pp. 1-7, 2024.

[22] M. Radhakrishnan, K. Ramamurthy, K.K. Choudhury, D. Won and T.A. Manoharan, "Performance Analysis of Deep Learning Models for Detection of Autism Spectrum Disorder from EEG Signals" *Traitement du Signal*, Vol 38, No. 3, pp. 1-6, 2021.

[23] X. Cao, W. Ye, E. Sizikova, X. Bai, M. Coffee, H. Zeng and J. Cao, "Vitasd: Robust Vision Transformer Baselines for Autism Spectrum Disorder Facial Diagnosis", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-5, 2023.

[24] X. Yu, J. Wang, Y. Zhao and Y. Gao, "Mix-ViT: Mixing Attentive Vision Transformer for Ultra-Fine-Grained Visual Categorization", *Pattern Recognition*, Vol. 135, pp. 1-7, 2023.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", *Proceedings of the International Conference on Computer Vision*, pp. 12-22, 2021.

[26] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang and J. Feng, "Deepvit: Towards Deeper Vision Transformer", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-6, 2021.

[27] W. Yuan, X. Zhang, J. Shi and J. Wang, "LiteST-Net: a Hybrid Model of Lite Swin Transformer and Convolution for Building Extraction from Remote Sensing Image", *Remote Sensing*, Vol. 5, No. 8, pp. 1-9, 2023.

[28] A. Kalaiselvi, S. Nagarathinam, T.D. Paul and M. Alagumeenaakshi, "Detection of Autism Spectrum Disorder using Transfer Learning", *Turkish Journal of Physiotherapy and Rehabilitation*, Vol. 32, No. 2, pp. 926-933, 2021.

[29] W. Yuan and W. Xu, "MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images based on Swin Transformer", *Remote Sensing*, Vol. 13, No. 23, pp. 1-6, 2021.

[30] H. Yang and D. Yang, "CSwin-PNet: A CNN-Swin Transformer Combined Pyramid Network for Breast Lesion Segmentation in Ultrasound Images", *Expert Systems with Applications*, Vol. 213, pp. 1-7, 2023.

[31] B. Juba and H.S. Le, "Precision-Recall Versus Accuracy and the Role of Large Data Sets", *Proceedings of International Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 4039-4048, 2019.

[32] W. Lu, C. Lan, C. Niu, W. Liu, L. Lyu, Q. Shi and S. Wang, "A CNN-Transformer Hybrid Model based on Swin Transformer for UAV Image Object Detection", *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 16, pp. 1211-1231, 2023.

[33] K.C. Raja and S. Kannimuthu, "Conditional Generative Adversarial Network Approach for Autism Prediction", *Computer Systems Science and Engineering*, Vol. 44, No. 1, pp. 1-7, 2023.