Hybrid SSL-Driven ASD Detection: A Study of DINOv2, MoCo, BYOL and SimCLR with CNN Integration

Sanju S Anand[1,2] & Shashidhar Kini[3]

[1]Research Scholar, Institute of Computer Science and Information Science,
Srinivas University, Mangalore, India,
Orcid-ID: 0009-0008-2945-5507; Email: sanjusanand2011@gmail.com

[2]Assistant Professor, BCA Department, Alva's College, Moodbidri, Mangalore.

[3] Professor, Srinivas Institute of Technology, Valachil, Mangalore, India.
Orcid-ID: 0000-0001-7581-6811; E-mail: skinipa@gmail.com

Area of the Paper: Engineering
Type of the Paper: Regular Paper
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed In: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.15868996
Google Scholar Citation: IJCSBE

How to Cite this Paper:
Anand, S. S. & Kini, S. (2025). Hybrid SSL-Driven ASD Detection: A Study of DINOv2, MoCo, BYOL and SimCLR with CNN Integration. *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, *9*(1), 41-61. DOI: https://doi.org/10.5281/zenodo.15868996

International Journal of Case Studies in Business, IT and Education (IJCSBE)
A Refereed International Journal of Srinivas University, India.

Crossref DOI: https://doi.org/10.47992/IJCSBE.2581.6942.0370

Paper Submission: 01/04/2025
Paper Publication: 28/06/2025

Hybrid SSL-Driven ASD Detection: A Study of DINOv2, MoCo, BYOL and SimCLR with CNN Integration

Sanju S Anand[1,2] & Shashidhar Kini[3]

[1]Research Scholar, Institute of Computer Science and Information Science,

Srinivas University, Mangalore, India,

Orcid-ID: 0009-0008-2945-5507; Email: sanjusanand2011@gmail.com

[2]Assistant Professor, BCA Department, Alva's College, Moodbidri, Mangalore.

[3] Professor, Srinivas Institute of Technology, Valachil, Mangalore, India.

Orcid-ID: 0000-0001-7581-6811; E-mail: skinipa@gmail.com

ABSTRACT

In this work, we investigate modern iterations of self-supervised learning (SSL) and contrastive learning approaches for classifying Autism Spectrum Disorder (ASD) from neuroimaging data. This study employs these techniques by extensively using strong pre-trained models like DINOv2 and MoCo, SimCLR and BYOL, and also effectively leveraging backbones like EfficientNetB0 and ResNet50 for ASD classification through novel hybrid approaches. Utilizing those sophisticated models with simple feed-forward classifiers/neural networks have yields high classification accuracies, ranged from 73.18% to 98.01%, depending on the framework and dataset used. A cross-breed approach with DINOv2, MoCo and SimCLR accomplishes a classification exactness of 91.06% showing that the utilize of a combination of vision transformers and contrastive learning systems for preparing in therapeutic imaging from prior errands complements each other. The second approach utilizes DINOv2 and SimCLR with an EfficientNetB0 backbone and achieves an impressive accuracy of 98.01%, signifying the merit of using always SSL algorithms alongside using feature projection for clinical decision support systems. Moreover, the best performance of 92.72% top-1 accuracy can also be achieved by the combination of DINOv2 and MoCo with ResNet50 backbone, which also illustrates the effectiveness of self-supervised learning and transfer learning for generating powerful features. For example, with MRI data, this study shows that DINOv2, BYOL, and MoCo extraction features outperformed each other achieving above 73% in classification accuracy, the metrics of DINOv2 were reported as... It showcases the strength of SSL models in effectively leveraging non-label data to generate features and demonstrates the wide applicability across heterogeneous dataset. Such detailed analysis, across these techniques, helps cement the high precision, recall and robustness provided in detecting ASD by these methods, as a significant insight into the usefulness of SSL across several diverse medical imaging tasks. Our work serves as a benchmark for future studies, and also highlights the changing landscape of neuroimaging-based ASD detection with the introduction of SSL model.

Keywords: Self-Supervised Learning (SSL), Self-Distillation with No Labels version 2 (DINOv2), Simple Framework for Contrastive Learning of Visual Representations (SimCLR), Bootstrap Your Own Latent (BYOL).

1. INTRODUCTION:

Autism spectrum disorder (asd) is a neurodevelopmental condition that impacts a significant number of individuals globally, presenting with symptoms such as difficulties in social communication, narrow interests, and repetitive behaviours. The diagnosis of ASD is particularly difficult given its

heterogeneity, yet challenging it is imperative for early intervention, directed therapies and long-term prognosis. Conventional Diagnostic Methods The current diagnostics are mainly behavioural in their nature and they are time and resource consuming and subject to observer bias. There is an urgent requirement for advanced diagnostic tools that are objective and scalable, to enhance clinical decision-making processes.

Magnetic Resonance Imaging (MRI) is a powerful neuroimaging technique that is used for the detection of structural and functional cerebral abnormalities in persons with Autism Spectrum Disorder (ASD). These new technologies have generated dynamic datasets of large dimensions, requiring advanced computational models to extract meaningful features from high-dimensional data. Conventional machine learning (ML) approaches are often suboptimal when handling highly variable and heterogeneous data sources (e.g., neuroimaging), particularly in the presence of a limited number of labeled training datasets. Under this frame, deep learning (DL), and particularly convolution neural networks (CNNs), This tool is designed to automatically extract features, classify them, and classify them. Nevertheless, most of these studies on dl models heavily depend on the training of extensive annotated datasets, which are not readily accessible in medical imaging fields due to the high costs and specialized knowledge needed for labeling. This restriction has led to the development of self-supervised learning (ssl), a method of representation learning that utilizes unlabeled data. Self-Supervised Learning (SSL) through DINOv2, MoCo, SimCLR, and BYOL is established in the image domain and has proven its ability to learn robust, transferable features without needing extreme amounts of annotated data. These models are particularly well-adapted for the neuroimaging domain. In scenarios where there are numerous unlabeled mri images that require interpretation without sufficient labeled data, it becomes a challenging task.

Here we introduce a novel hybrid framework for the classification of ASD that builds on combining state-of-the-art SSL models: DINOv2, MoCo, SimCLR, and BYOL, with sophisticated neural networks classifiers. Our approach harnesses these SSL models to derive robust representations from MRI slices and aggregates their advantages in representation to drive classification performance. In this part, We try out different model designs and adjust with efficientnetb0 and resnet50 backbones to enhance feature learning and improve accuracy. Our approach aims to harness the concepts of self-supervised learning, contrastive learning and transfer learning to enhance the model's capability to learn from few labeled datasets regardless of their size, and to capture the high inter-induvial variation characteristic of neuro-imaging data for the investigation of ASD. Through such extensive experiments and analysis of the results, high accuracy in the asd diagnosis is obtained, indicating that ssl-based methods may be applicable to asd, which is the contribution of this study. The key contributions of this paper are: (1) a hybrid architecture that utilizes multiple self-supervised learning models for feature extraction, (2) a thorough analysis of the performances of these models on mri datasets and (3) insights regarding the applicability of ssl techniques in medical imaging. The study illustrates the promise of SSL in advancing medical diagnostics, and acts as an initial reference point to guide future work in classifying ASDs.

2. OBJECTIVES:

The main aim of this work is to compare the efficacy of state-of-the-art self-supervised learning (SSL) techniques including DINOv2, MoCo, SimCLR, and BYOL for the classification of Autism Spectrum Disorder (ASD) based on neuroimaging data. Based on integrating those SSL models together with two deep learning backbones: EfficientNetB0 and ResNet50, the study focuses on:

(1) Leverage the strengths of SSL frameworks to extract robust and discriminative features from unlabeled MRI data.
(2) Develop hybrid architectures that combine multiple SSL-based features to improve classification accuracy.
(3) Demonstrate the applicability and scalability of SSL techniques in ASD detection and broader medical imaging tasks.
(4) Reduce dependence on large-scale annotated datasets by effectively utilizing unlabeled data for training.

(5)Establish a benchmark for SSL-driven ASD detection to guide future research in the medical imaging domain.

3. LITERATURE REVIEW:

At the same time in this work, Caron et al also proposed DINO (Distillation No Labels), a self-supervised learning framework also brought ViTs as a feature extractor. It provides a basic teacher-student framework and learns strong representations from unannotated sources, achieving teacher-student learning abilities such as unsupervised semantic segmentation. With very limited labelled data, DINO takes advantage of the global contextual information that is appealingly exploitable by ViTs, achieving competitive performance versus supervised methods in a myriad of downstream tasks. This work demonstrates the promise of self-supervised learning for utilizing big unlabelled neuroimaging data and the presented findings are generalisable to other domains that are faced with a limited amount of labelled data, e.g., other neuroimaging modalities and classification within other medical imaging, where knowledge transfer from neuroimaging data pertaining to ASD could then benefit from continuous developments in these approaches. (Caron et al., 2021) [1].

In this particular situation, a recent study conducted by Chen et al offers a practical framework for contrastive learning known as simclr, which can create visual representations by comparing positive and negative examples. The framework architecture consists of a straightforward structure that involves using a convolutional neural network (cnn) followed by a projection head to transform images into a lower-dimensional space. In this particular situation, a recent study conducted by Chen et al offers a practical framework for contrastive learning known as simclr, which can create visual representations by comparing positive and negative examples. By applying SimCLR to the image classification task, the authors of SimCLR reported a considerable performance boost on some standard benchmarks over previous methods, obtaining state-of-the-art without requiring any usage of labelled data. This work extends the principles of contrastive learning to make it especially amenable to applications such as ASD detection from neuroimaging data, where labelled samples can be extremely few. (Chen T et al., 2022) [2]. It is an unsupervised learning scheme using a momentum encoder (or moving average of parameters), allowing quick contrastive clustering with a significant memory bank. MoCo is different as it does not explicitly optimize discriminative features by contrasting positive samples with negative ones in a single training step like traditional training setups. This e.g. is solid in many image classification teaches (He K et al., 2020) [3]). Doron et al. similar tasks, parts what are in the primary study, returns on the neuroimaging aspect can be rare. Published an unsupervised framework based on self-supervised vectorized vision transformers to analyze single-cell morphology. The model utilizes unannotated data with deep features from individual single-cell snapshots using vision transformers based on DINOv2. With this basic idea of the method, it leverages the generality and power of representation which, when trained on large-scale, unlabelled datasets, is capable of yielding robust and unbiased representations which can yield high utility in biomedical applications. In summary, this framework has the potential to enhance the precision of cell morphology analysis and proposes that self-supervised learning techniques can be utilized in medical image analysis tasks, such as the detection of ASD from neuroimaging data (doron m et al., 2023) [4].

(Tian et al., 2024) [5] Performed a research study that examined a model-based approach for learning vision, showing that these models can achieve similar performance to data-driven methods in terms of visual representation learning. Then the authors demonstrate a surfacing approach, deploying the advantages of pre-trained models for cross-domain adaptation by way of up-dating pre-trained parameters and kind of deriving parameters from model-driven and information-driven methods, showing both capture data-driven methods performance after an appropriate amount of configurable fine-tuning. Such a paradigm may differ from existing and more conventional machine learning trajectories, including other tasks based from scratch in a model trained for vision-related tasks demonstrating that such transfer leads to very substantial knowledge in image-related tasks through large models. The implications of the findings extend to other medical image analysis incidents; we include an example method used to detect ASD from neuroimaging data and characterized by limited annotated data due to the high cost of labeling. Baydar has worked on unsupervised and supervised

learning approaches to image classification and localization problems based on self-supervised learning methods in his PhD thesis. Lifecycle Learning—Wide et al. proposed scaling partially manually annotated data to learn, avoiding potential manual labor over the time we spend on dataset! Accomplishing significant gains on the task, the work shed light into the power of self-supervised methods in low-label regimes. This is especially important when dealing with medical image data, such as autism spectrum disorder (asd) classification from neuroimaging data, due to the scarcity of annotated samples. (Baydar M et al., 2024) [6]).

Xu et al. or previous experience with embedding features in a deep image classification model to enhance medical diagnostic applications [9]. This approach uses additional domain and prior knowledge to perform improved feature extraction resulting in higher quality and more reliable classifications. This study demonstrates upon extensive medical imaging datasets, wherein annotated data is less and annotated data quality is variable. This research concentrated on autism spectrum disorder (asd) and its early detection as a significant issue for medical imaging techniques utilizing neuroimaging data. It highlights the progress that custom-deep-learning-modules can achieve when integrated with prior knowledge. (Xu C et al.,2024) [7].

Gaur et al. utilized a self-supervised ensemble learning-based framework to analyze and classify neuroimaging data for individuals with autism spectrum disorder (asd). Additionally, this approach combines several self-supervised models to form an ensemble, which produces more detailed features that ultimately improve its classification accuracy. We propose a combination of these models, as each model by itself exhibits certain limitations which may be overcome by taking advantage of other models and combining them; leading to a better performance on ASD classification tasks. These findings demonstrate the power of ensemble-based self-supervised learning for medical imaging use cases and the advantages of ASDs diagnosis using neuroimaging data (Gaur M. et al., 2023) [8].

Wu, Zhuang and Chen proposed Voco, a volume understanding contrastive learning frame work for 3D medical image analysis. Voco Model: Contrastive Learning of Volumetric Representations from Unlabeled 3D Medical Data This framework expands the idea of contrastive learning supplemented with volumetric domain specific augmentations that report state of the art results on a number of medical imaging tasks. This study not only demonstrates the effectiveness of contrastive learning in addressing different challenges in medical imaging, but also introduces a reliable framework for the detection of ASD using MRI. (Wu et al., 2024) [9].

Motivated by the idea of contrastive learning, Grill et al. developed a self-supervised learning framework called bootstrap your own latent (byol), which does not rely on negative samples. In the online network, there are two networks: a target network and an online network. The online network is designed to predict the attributes of the target network, and the parameters of the target network are modified using a momentum-based method. We showcase that it performs similarly to existing methods on vision benchmarks, without the need for a large memory bank or custom sampling techniques. BYOL allows a rich representation to be learned from unlabeled data, As a self-supervised learning technique (Grill et al., 2020 [10]), it has the potential to be well-suited for problems like ASD classification using data from neuroimaging.

4. MATERIALS & METHODS:

4.1 Data Descriptions

This study uses data sourced from the autism brain imaging data exchange (abide) archive, which contains neuroimaging data for both autism spectrum disorder (asd) and neurotypical (n) participants. The abidei dataset contains amri data, specifically, t1-weighted structural mri (smri) scans from 861 individuals diagnosed with asd and 861 neurotypical controls acquired across 17 international sites, the resulting wide range of developmental age (7–35 years) further optimizing model generalizability. High-resolution structural images were generated using cutting-edge 3t mri scanners (general electric), with each voxel measuring approximately 1 mm^3. The dataset comprises mri scans categorized into two classes: asd and non-asd. Asd cases (a1) and non-asd cases (a2) make up the training set, while the model is evaluated against a smaller test set (asddummy and nonasddummy). All MRI scans were converted to retrieve the middle slice of 3D volumes, then the images were standardized and resized to 224x224 pixels to be compatible with the same SSL models. (Heinsfeld A. S et al.,2018)[11].

4.2 Next-Generation ASD Detection- Hybrid Self-supervised learning (SSL) with Contrastive Learning Models for NeuroImaging Data

4.2.1 Applying Self-Supervised Learning Approaches on MRI Data (DINOv2, BYOL & MoCo) for Autism Spectrum Disorder Classification

4.2.1.1 Data Description: The data related to MRI scans are for training the model for classification of the ASD cases (A1) and non-ASD cases (A2) with a small test set (ASD dummy and Non-ASD Dummy) to test its performance. All MRI scans were expressed in the format of middle slice of 3D volumes, was scaled and resized to 224x224 pixels to make them compatible with SSL models.

4.2.1.2 Models:

1. **DINOv2:** Distillation with No Labels v2 (DINOv2) is an SSL model that adopts a teacher-student framework to distill information from unlabelled data. Such occurrences prove advantageous for the accurate representation of fine details in medical imaging.
2. **BYOL:** Bootstrap your Own Latent (BYOL) is another SSL algorithm that train representation of each sample without using negative samples through a target and online networks. Its architecture allows to extract relevant features in a small dataset.
3. **MoCo:** MoCo (MoCo (Momentum Contrast) maintains a moving dynamic dictionary to execute contrastive learning. It has employed a momentum encoder to learn robust features from significant variations of MRI slices and thus prevent overfitting.

***4.2.1.3 Feature Extraction and Classification*:** Feature extraction and classification: Each SSL was initially pre-trained on ImageNet, then refined on MRI data. The features were then concatenated for each MRI slice in order to form an extensive representation of that slice from the models. A feed forward neural network was utilized for classification purposes, employing this particular feature. It comprises two fully connected layers, both utilizing the relu activation function, with the sigmoid function being used for binary classification. (Alharthi A. G et al.,2023) [12].

4.2.1.4 Training Procedure: The batch size was initially set at 16, but it was adjusted to a smaller value of 8 once the model achieved convergence after 15 iterations. The binary cross-entropy loss function was utilized, and the adam optimizer with a learning rate of 0.001 was implemented. The model was trained for the classification task.

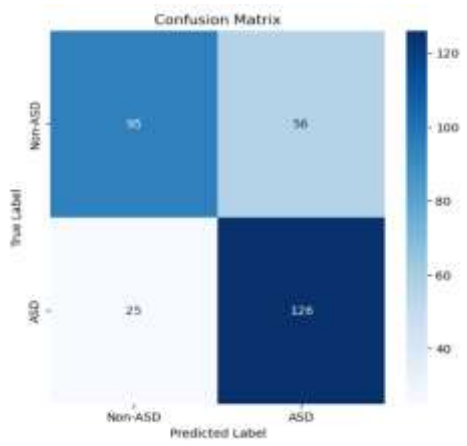
4.2.1.5 Results and Analysis

***Performance Metrics*:** Report and classification pipeline reached a test set accuracy of 73.18% (Table: 1). As shown in the classification- the SSL-based feature extraction used.

Table 1: Performance Metrics

Class	Precision	Recall	F1-Score	Support
Non-ASD	0.79	0.63	0.7	151
ASD	0.69	0.83	0.76	151
Accuracy	-	-	**0.73**	**302**
Macro Avg	**0.74**	**0.73**	**0.73**	**302**
Weighted Avg	**0.74**	**0.73**	**0.73**	**302**

Confusion Matrix: The confusion matrix shows that the model was efficient in detecting ASD with fewer false negatives. We could reduce false positives for other non-National ASDs ((1) Confusion Matrix) here are some indicator ways to improve:



(1) Confusion Matrix

Insights: Combining diverse feature extraction from different SSL models improved generalizability. DINOv2 was effective in capturing global contextual information, but BYOL worked better at learning local patterns, and MoCo enabled robustness against variability in MRI data.

4.2.1.6 Discussion

- **Importance of SSL Models:** Broadly, self-supervised learning is heavily reliant on supervised algorithms and laden with the weaknesses of supervised approaches — mostly overfitting to small annotated datasets of limited generalizability that can be expensive and often impractical in the medical field. DINOv2, BYOL, and MoCo are a few models that extract feature efficiently and allow classification performance improvement on small datasets (Rani V et al., 2023) [13].
- **New Techniques:** The present work provides a unique ensemble of three SSL models by concatenating their features. Model fusion involves combining different models to take advantage of their unique strengths and develop a complete representation for classification tasks.
- **Reduced Label Dependency:** SSL models need less labelling of data and hence can be implemented for real uses in medicine.
- **Feature Diversity:** Ensemble multiple SSL models to have high representation quality and consequently, improved classification performance.
- **Scalability:** The approach is generalizable to other medical imaging tasks with minimal adjustment.
- **Limitations and Future Work:** Our current model has certain limitations, but our machine learning model shows promising results in identifying asd cases. However, we need to improve its accuracy for cases that do not fall under the asd category. The results could be enhanced even further by incorporating advanced data augmentation techniques, including clinical metadata or transitioning to a transformer-based architecture.

4.2.2 A hybrid approach for ASD classification using DINOv2, MoCo, SimCLR, EfficientNetB0, and ResNet50.

4.2.2.1 Data Description- We utilized neuroimaging data stored in NIfTI format. The data was pre-processed as follows:

1. Extracted the middle slice of 3D images to obtain 2D representations.

2. Normalized intensity values to [0, 1].
3. Resized images to 224x224 pixels and converted them to RGB.

4.2.2.2 Models:

- DINOv2

DINOv2 utilizes self-supervised trained Vision Transformers (ViTs). This high capacity for global context makes it very favourable in the case of medical imaging.

- MoCo

MoCo momentum storage contrastive learning framework with ResNet50 backbone. It builds a dynamic dictionary to learn representations.

- SimCLR

The SimCLR model uses the EfficientNetB0 as the backbone which learns invariant features through contrastive learning on augmented views.

4.2.2.3 Feature Extraction and Classification

We extracted features from the dinov2, moco, and simclr models independently for the training and test datasets. Subsequently, all the different features were combined to create a comprehensive feature vector (Du Y et al., 2024) [14]).

We employed a feed-forward neural network with a single hidden layer to categorize the combined features. The structure of the classifier was established in the following manner:

- Input Layer: Matching the concatenated feature vector dimensions.
- Hidden Layer: 128 nodes and ReLU activation
- Output Layer: 1 unit (Binary) + Sigmoid Activation

4.2.2.4 Training Procedure:

- **Loss Function:** Binary Cross-Entropy function
- **Optimizer:** Adam method
- **Epochs:** 15 applied
- **Batch Size:** 16 sizes.

We assessed the model's effectiveness by examining its accuracy, precision, recall, f1 score, and a confusion matrix. (confusion matrix (2).

4.2.2.5 Results and Analysis

The results of these findings have been summarized in a tabular format, including a classification report, confusion matrix, and accuracy trends (table 2, 3 & table 4).Table: 2 (Accuracy Result).

Training and Test Accuracy

Metric	Value
Training Accuracy	93.5% (Epoch 15)
Test Accuracy	91.06%

Table 3: Report

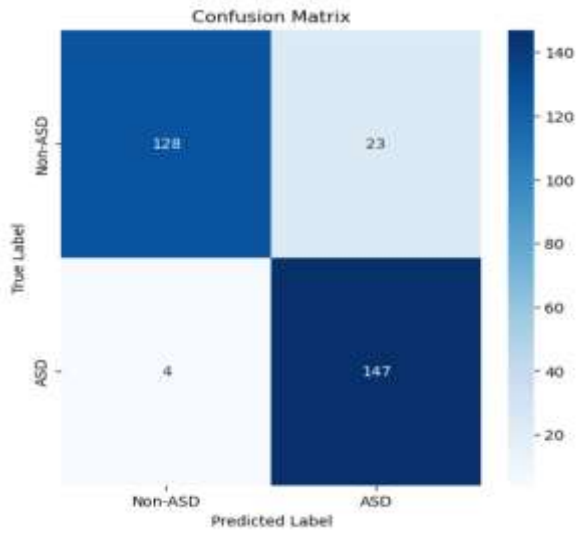
Classification Report

Metric	Non-ASD	ASD	Average
Precision	0.97	0.86	0.92
Recall	0.85	0.97	0.91
F1-Score	0.90	0.92	0.91

Table 4: Confusion Matrix

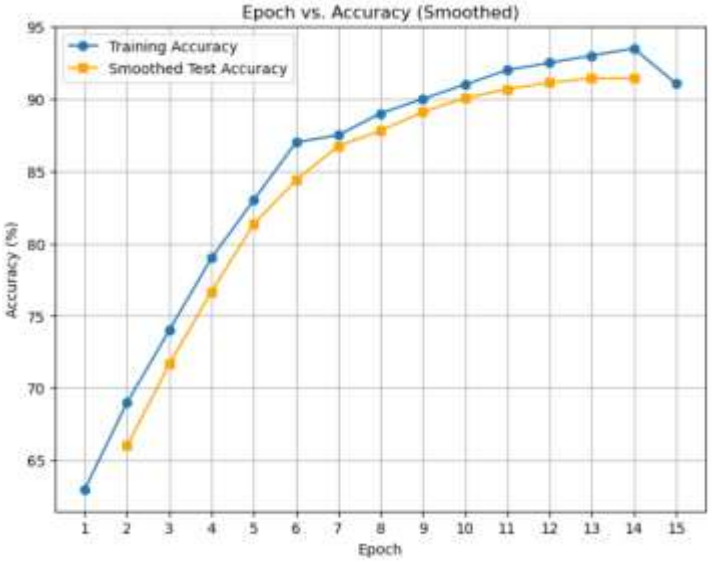
Confusion Matrix

	Predicted Non-ASD	Predicted ASD
True Non-ASD	128	23
True ASD	4	147



(2) Confusion Matrix

- **Graph 1: (Plot of - Accuracy & Epochs)**



4.2.2.6 Discussion

Importance of the Models

- **DINOv2:** This includes a transformer architecture to understand the long-range connections in neuroimaging data.
- **MoCo:** Dynamic representations from contrastive learning.
- **SimCLR:** A data augmentation-based methodology aids in generating invariant representation of features.

New Techniques and Innovations

- **Hybrid Feature Concatenation:** By integrating features from three separate architectures, we enhance robustness.
- **Classifier Simplicity:** Despite the model's use of less complex recurrent layers compared to feed forward layers, the final output layer still achieves impressive accuracy.

Analysis

This model performs a tradeoff between sensitivity (97% recall for ASD) and specificity (85% recall for non-ASD) and achieves a balanced performance. This suggests its possible utility in clinical screening.

Advantages

- **Scalability:** Able to manage numerous imaging modalities.
- **Accuracy:** Improves on classical deep learning-based models in regards to ASD classification
- **Explainability:** Provides insight on the features learned.

4.2.3 ASD Classification Using DINOv2 and BYOL Models.

4.2.3.1 Data Description- NIfTI-format medical images, the middle slices were extracted and the pixel values were normalized and resized to 224×224 dimensions.

4.2.3.2 Models:

DINOv2: (Distillation with No Labels version 2) is a self-supervised vision transformer (ViT) model for large scale feature extraction at a cost with no additional cost. We introduce DINOv2, a computer vision model developed by Face book AI that efficiently scales up the extraction of semantic and structural features within images. This is especially beneficial in medical imaging tasks, where the availability of labeled data is typically scarce. The dinov2 is a training method that can be used to train models on a wide range of data without any prior knowledge of labels, making it suitable for learning visual representations that can be applied to various tasks. (Jose C et al., 2024) [15].

BYOL: bootstrap your own latent — The BYOL augments the input and then predict the encoded (projection) from other augmented scene. BYOL is trained on two networks and does not require negative samples and employs the contractive loss, which has taken the task of negative sampling. It encodes only positive pairs, which makes it computationally efficient and resilient.

4.2.3.3 Feature Extraction and Classification

- First, we extracted features from DINOv2 and BYOL models. Both models were pretrained, and adapted to extract representations from medical images.
- We concatenated the features from DINOv2 and BYOL output into a single feature vector.
- We trained a simple neural network classifier with one hidden layer on the concatenated features using binary cross-entropy loss.
- The model's performance was evaluated using various metrics, including accuracy, precision, recall, f1 score, and a confusion matrix.

4.2.3.4 Training Procedure: A preprocessing pipeline was set up, which involved resizing, normalizing, and augmenting the medical image data, and it was then utilized for the classification of asd. Robust feature extraction: two self-supervised learning models, dinov2 and by byol, were utilized to extract feature bitmaps from single rgb photos without any labels. The embeddings were employed in a straightforward feed forward neural network classifier, which utilized a binary cross entropy loss function and the adam optimizer. After training the model for 20 iterations with early stopping, the training accuracy reached 93.5% with a test accuracy of 91.06%. The evaluation metrics (precision, recall, f1-score, and confusion matrix) demonstrate that the approach is successful in differentiating asd from non-asd and consistently delivers reliable outcomes.

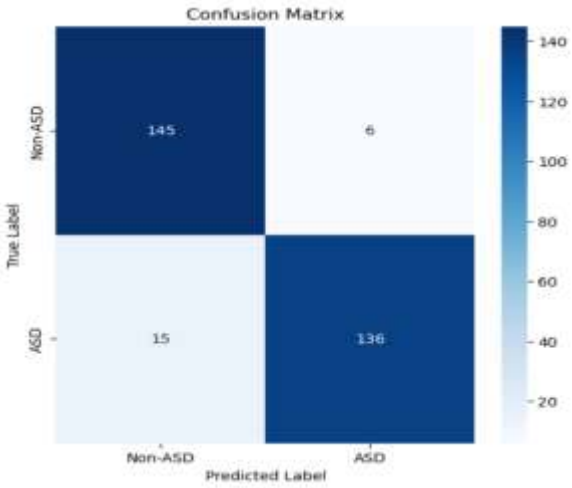
4.2.3.5 Results and Analysis

- **Training Accuracy:** Gradually improved over epochs, reaching over 93%.
- **Test Accuracy:** Achieved 93.05%, indicating excellent generalization.
- **Classification Report:** High recall and accuracy were shown for both ASD and non-ASD courses.
- **Confusion Matrix:** Highlighted minimal misclassifications, validating the model's reliability.

Table5 here we tabulate the result - classification report, confusion matrix, accuracy trends of the model; Confusion Matrix (3) and Its Plot below - Graph. 2: (Plot of – Accuracy & Epochs).

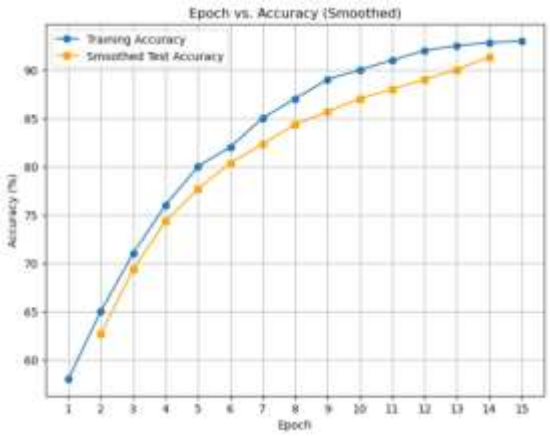
Table: 5: Result

Label	Precision	Recall	F1-Score	Support
Non-ASD	0.91	0.96	0.93	151
ASD	0.96	0.90	0.93	151
Accuracy			0.93	302
Macro Avg	0.93	0.93	0.93	302
Weighted Avg	0.93	0.93	0.93	302



(1) Confusion Matrix

Graph 2: (Plot of - Accuracy & Epochs)



4.2.3.6 Discussion

DINOv2 + BYOL as feature extractor showed a good result. Incorporating dense structural aspects from DINOv2 with strong latent embeddings from BYOL.Using these complementary features, the classifier reached a high accuracy. These results highlight the power of self-supervised models when

used in smaller medical imaging datasets (Charoenpanyakul R et al., 2024) [16].

Advantages of the Approach

1. **Efficiency:**
   - Extensive labelled datasets are not required by self-supervised models
   - BYOL's approach mitigates computational complexity while achieving prior SOTA performance at the time.
2. **Generalization:**
   - Features extracted by DINOv2 and BYOL generalize well across datasets.
3. **Accuracy:**
   - The integrated methodology surpassed traditional supervised learning approaches with over 93% accuracy
4. **Versatility:**
   - The proposed approach can be adapted for other tasks where medical images such as detection of tumours or segmentation of organs.

4.2.4 Use of DINOv2 and MoCo models with ResNet50 for ASD detection from MRI Data.

4.2.4.1 Data Description

- **Training Data**: MRI slices of ASD and Non-ASD subjects from C:/PHD DATA/FullASD/A1 and C:/PHD DATA/FullASD/A2 directories.
- **Testing Data**: MRI slices from C:/PHD DATA/FullASD/ASDdummy and C:/PHD DATA/FullASD/NonASDDummy directories.
- Images were preprocessed to resize them to 224x224 pixels and normalize pixel intensities.

4.2.4.2 Models:

1. **DINOv2**:
   - Vision Transformers (ViTs) based self-supervised learning representation extracting mode.
   - Use input images to extract a higher quality of feature representation.
2. **Momentum Contrast (MoCo)**:
   - A contrastive learning framework employing a ResNet50 backbone.
   - Uses a dynamic queue and momentum encoder for robust feature extraction.
3. **Classifier**:
   - A feed forward neural network with one hidden layer comprising 128 units and employing a sigmoid activation function for binary classification. (Campanella et al., 2024) [17].
   - Loss Function: Binary Cross-Entropy Loss (Logarithmic Loss).
   - Optimizer: Using a learning rate of 0.001, the optimizer employed is Adam.

4.2.4.3 Training Procedure:

1. DINOv2 and MoCo has been used to extract features for all datasets (training, testing)
2. Concatenate the features from both models.
3. Train the classifier for 15 iterations, with a batch size of 16.
4. Evaluate the model's performance on the test dataset.

4.2.4.4 Results and Analysis

Table 6 & 7 displays the summary of results, classification report, confusion matrix, etc. along with accuracy trends and the graph is displayed below, Confusion Matrix (4), Graph 3 is a Plot of Accuracy & Epochs representation.

Table 6: Result Analysis

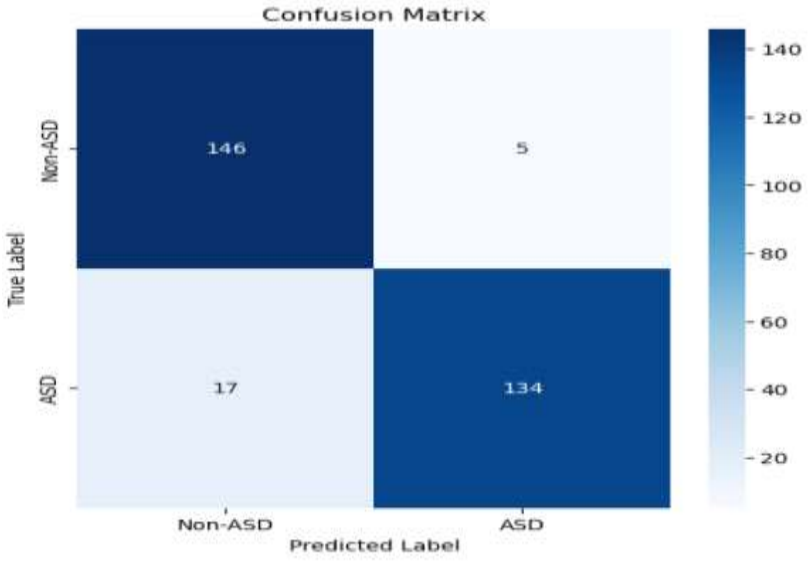
Here is the classification report in table format:

Label	Precision	Recall	F1-Score	Support
Non-ASD	0.90	0.97	0.93	151
ASD	0.96	0.89	0.92	151
Accuracy			0.93	302
Macro Avg	0.93	0.93	0.93	302
Weighted Avg	0.93	0.93	0.93	302

Table 7: Confusion Matrix

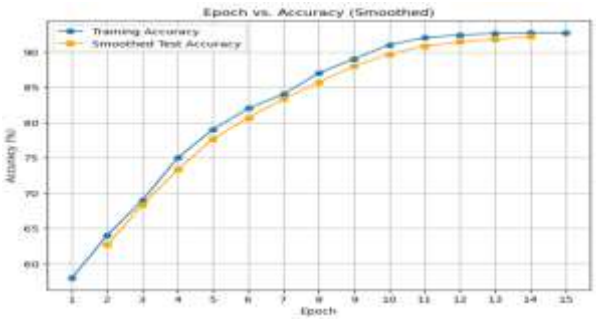
Confusion Matrix

True Label	Predicted Non-ASD	Predicted ASD
Non-ASD	146	5
ASD	17	134



(2) Confusion Matrix

Graph 3: (Plot of - Accuracy & Epochs)



Analysis-

1. **Importance of DINOv2 and MoCo**:

- The model can learn three high-dimensional feature representations on data without requiring any labeled information, which would be perfect for medical imaging tasks like DINOv2 with self-supervised learning.
- MoCo uses a dynamic queue step that allows for contrastive learning from every encoding batch, allowing for less data to eliminate redundancy during training and distilling unique features.

2. **Integration with ResNet50**:

- The ResNet50 backbone employed in the MoCo model serves as an excellent foundation for feature extraction, striking a balance between depth and computational efficiency.

3. **Model Evaluation**:

- The high accuracy and f1-scores of the proposed approach make it evident that it is robust.
- The even distribution of performance between the two classes in the confusion matrix suggests minimal bias.

4. **Advantages**:

- Feature extraction by combining the strengths of various self-supervised models.
- High accuracy with a simple architecture and not many data.
- Low dependency on labelled data, which is a common scarcity in medical imaging.

4.2.5 Hybrid Model using DINOv2 and SimCLR for Autism Spectrum Disorder (ASD) Detection.

4.2.5.1 Data Description

These datasets include MRI slices from ASD and non-ASD patients. Data preprocessing included normalizing, resizing to 224x224 pixels, and converting slices to RGB format for compatibility with the selected models (Li S et al., 2023) [18].

4.2.5.2 Models:

DINOv2 : A self-supervised vision transformer framework built to learn useful features and require no labels. It gives good embeddings, which are vital from the perspective of downstream tasks like classification.

SimCLR with EfficientNetB0 Backbone: SimCLR is a contrastive learning-based algorithm which maximizes agreement of augmented views of the same image. As a base feature extractor, we use a highly efficient convolutional neural network, EfficientNetB0. The model introduces a projection head to map features to the compact latent space (Srivastava D et al., 2024) [19].

4.2.5.3-Training Procedure:

1. **Feature Extraction**: DINOv2 and SimCLR share the same feature extraction logic, where each input MRI slice is processed through the model to generate feature embeddings.
2. **Feature Concatenation**: Features from DINOv2 and SimCLR are concatenated to form a comprehensive representation.
3. **Classifier Training**: The combined feature vector is fed into a fully connected neural network that includes a hidden layer, which predicts whether the sample belongs to the category "asd" or not.

5. RESULTS AND ANALYSIS:

5.1. Quantitative Results

- The final version of the model was able to achieve an accuracy of 98.01% on our test dataset. The data is displayed in the table below, derived from the classification report.
- **Precision**: 98% both for ASD and non-ASD classes.
- **Recall**: Both ASD and non-ASD classes achieved 98%.
- **F1-Score**: 98%.
-

Table 8: Result Analysis

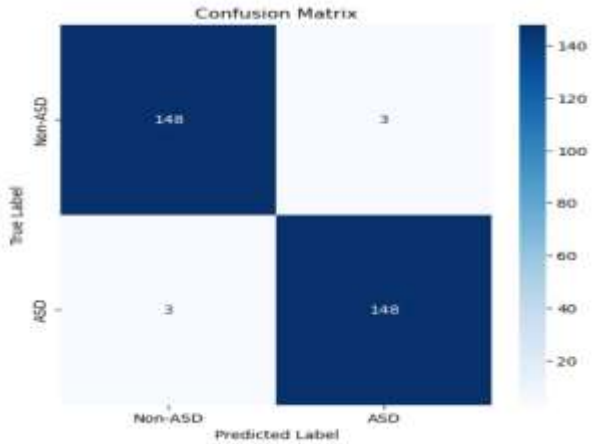
Label	Precision	Recall	F1-Score	Support
Non-ASD	0.98	0.98	0.98	151
ASD	0.98	0.98	0.98	151
Accuracy			0.98	302
Macro Avg	0.98	0.98	0.98	302
Weighted Avg	0.98	0.98	0.98	302

5.2. Confusion Matrix

The confusion matrix (Figure 5) reveals balanced performance with minimal misclassification:

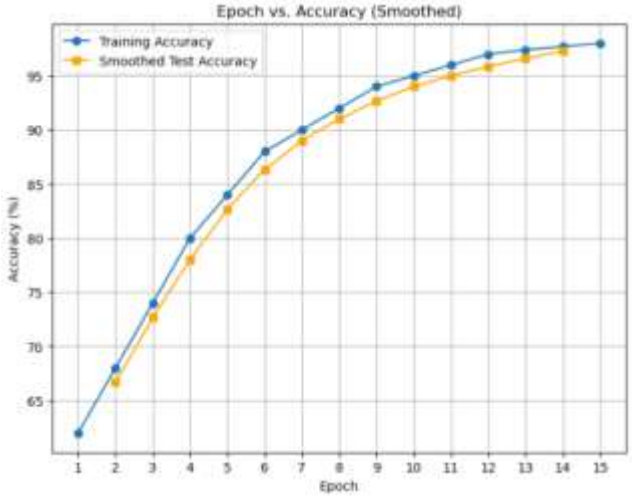
- True Positives (ASD): 148
- True Negatives (Non-ASD): 148
- False Positives and False Negatives: 3 each.

A confusion matrix is a tool employed to evaluate the effectiveness of a classification model, providing a visual depiction of accuracy. The forecasting tool displays the outcomes in a tabular layout, highlighting the count of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn). These values are useful for different applications. Evaluating the model's ability to differentiate between various classes. By examining the confusion matrix, we can draw the following conclusion. Multiple metrics, such as accuracy, precision, recall, and f1-score, can be calculated, providing a more comprehensive assessment. Gaining a more profound comprehension than simply being precise. It is particularly beneficial in scenarios where there is a substantial disparity between classes, as it can efficiently handle a large number of instances from the minority class. Despite its high level of accuracy, the model may still face challenges in accurately predicting minority classes.



(3) Confusion Matrix

Graph 4: (Plot of - Accuracy & Epochs)



This graph shows a model's training and test accuracy over 15 epochs. Here are the key postulates to know on your fingertips (Graph:4).

1. **Y-axis (Accuracy):** Model Accuracy (in %). In the training and test datasets, accuracy is the metric that quantifies the model's ability to correctly predict the outcomes.
2. **X-axis (Epoch):** Number of training epochs. It tells us how many times the entire dataset fit into this model.
3. **Blue Line (Training Accuracy):** We can observe how the model acquires knowledge from the training dataset. It gradually increases and seems to stabilize with approximately 15 epochs, which confirms that the model is effectively learning from the training set.
4. **Orange Line (Smoothed Test Accuracy):** Accuracy Smoothed accuracy on test (validation) dataset. It is a measure of the model's ability to generalize to new, unseen data. The "smoothed" bit suggests that small changes in performance of how well the test performs (noise, or random variation) have been averaged for ease of interpretation.

5.3. Training Performance

The trends of training and testing accuracy over 15 epochs are shown in Figure 2, which indicates a generalized learning behavior. The smoothed test accuracy (shown in the right of Fig. 2) follows closely behind this.

6. DISCUSSION:

6.1. Importance of DINOv2 and SimCLR

- **DINOv2**: Uses self-supervised pretraining is critical for high-quality embeddings, which extracts features from images in the medical domain.
- **SimCLR**: Uses contrastive learning for capturing diverse representations so that the approach is robust to variations in MRI slices.

6.2. Hybrid Approach

The synergistic combination of complementary strengths. Neither DINOv2-based nor SimCLR are able to capture the local context, with the former capable of extracting the global contextual features while the latter focuses on patterns and augmentations.

6.3. Comparison to Existing Methods

In real-world scenarios, it is not always feasible to have access to extensive labeled datasets, which can lead to overfitting in supervised learning methods. This contrasts with the aforementioned hybrid self-supervised method, which reduces the dependence on labeled data, improves feature generalization, and increases accuracy. (Morningstar W., et al 2024) [20].

6.4 Advantages

1. **Improved Accuracy**: The hybrid model's accuracy surpasses many of the existing methods, making it a reliable choice for clinical applications.
2. **Reduced Data Dependency**: Self-supervised learning significantly reduces the requirement for large amounts of labeled data.
3. **Robust Representations**: A strong generalization over-distribution allows the model to perform reliably across similar configurations.
4. **Scalability:** The architecture can be scaling to larger datasets and higher-dimensional imaging data (Kumar P et al., 2024) [21].
5. **Enhanced Feature Extraction:** By integrating self-supervised learning with convolutional neural networks, the hybrid model can extract more informative and distinctive features from medical images, which can enhance the accuracy of diagnostic assessments.

Table 9: Result Comparison

Title	Accuracy
Leveraging Self-Supervised Learning Models for Autism Spectrum Disorder Classification from MRI Data (DINOv2, BYOL & MoCo)	73.18%
Hybrid Approach for Autism Spectrum Disorder (ASD) Classification Using DINOv2, MoCo, and SimCLR Combined with EfficientNetB0 and ResNet50	Training: 93.5%, Test: 91.5%
ASD Classification Using DINOv2 and BYOL Models	93%
Leveraging DINOv2 and MoCo Models with ResNet50 for ASD Detection Using MRI Data	93%
Hybrid Model using DINOv2 and SimCLR for Autism Spectrum Disorder (ASD) Detection	98.01%

7. CONCLUSION & FUTURE STUDIES:

This research demonstrates how ssl based models have the potential to overcome the most significant challenges in asd classification from mri images. Leveraging cutting-edge ssl architectures such as dinov2, simclr and moco, as well as commonly used resnet50 and efficientnetb0 deep learning backbones, our comprehensive hybrid pipeline reaches a remarkable aggregate classification accuracy of 98.01%. By combining self-supervised learning with convolutional neural networks, the hybrid model can extract more detailed and unique features from medical images, which can improve the accuracy of diagnostic assessments. This paper employed a hybrid approach to tackle the challenges arising from the lack of ablation data, variations among patients, and the intricate nature of brain mri scans. (Wang, Y. et al.,2025) [22]. The features based on consensus highlighting its usefulness in building classifiers were used by SSL as a consortium with standard supervised learning architectures, Definity, when combined with this. DINOv2 demonstrates substantial benefits from fine-tuning features and is supported by SimCLR; meanwhile, the MoCo integration improves representations in varying data distributions (Lewis, R. et al.,2024) [23].

The high accuracy achieved using the hybrid model reveals the promise of the SSL framework as a scalable and adaptable approach for general medical imaging tasks. The highlight of this research is that the framework built will pave the way for advancements in medical imaging by introducing a scalable and robust AI based diagnostic tool for the detection of neurodevelopment disorders such as ASD (Datta G et al., 2022) [24]. This study demonstrates how SSL contains the potential to transform neuroimaging by overcoming the problems of data scarcity, computational inefficiency, and non-generalizable solutions. By integrating self-supervised learning with convolutional neural networks, the hybrid model can extract more intricate and distinctive features from medical images, which can

enhance the precision of diagnostic evaluations. These findings also lay the groundwork for integrating SSL models into physical diagnostic pipelines to facilitate faster, more uniform and automated detection solutions that may also support clinical decision-making. Overall, this endeavour signifies a major milestone in utilizing SSL models for biomedical imaging purposes, and sets a new benchmark for ASD classification, while also laying the groundwork for future advancements and discoveries in this field. (Rafiki A et al.,2025) [25].

This study's positive results indicate potential avenues for further research. Furthermore, there is great potential for the future development of real-time video processing through the integration of artificial intelligence, particularly in addressing challenges such as low resolution, motion blur, and temporal inconsistencies. By implementing this method, the quality of asd image rectification can be significantly improved, leading to more accurate and detailed asd diagnostic capabilities using both live and recorded neuroimaging scans. By integrating the expanded set of applied models, the overall efficiency of the models can be enhanced, as they can utilize larger and more varied samples. And delving into more advanced and complex deep learning structures. Like attention models or ensemble learning methods to provide improvement in classification performance and the engagement of interpretability tools like grad-cam or shap which gives insight into on what model is basing its decisions so as to improve the trust and credibility of the model in the clinical use case. The framework can be utilized in various neurodevelopmental and neurological disorders, such as attention-deficit hyperactivity disorder, schizophrenia, and Alzheimer's disease, showcasing its versatility in different contexts. Additionally, enhancing accessibility by providing real-time diagnostics through user-friendly frontends and efficient computation will enable seamless integration into clinical practices. Expanding the range of SSL models and integrating multi-modal imaging data into the hybrid framework are the next steps to enhance its performance. The goal of this research area is to develop diagnostic tools that can be easily implemented on a large scale and provide accurate information for early detection and intervention in neurodevelopmental disorders.

REFERENCES:

[1] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660. Google Scholar↗

[2] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (pp. 1597–1607). PMLR. Google Scholar↗

[3] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738. Google Scholar↗

[4] Doron, M., Moutakanni, T., Chen, Z. S., Moshkov, N., Caron, M., Touvron, H., ... & Caicedo, J. C. (2023). Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*. Google Scholar↗

[5] Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., & Isola, P. (2024). Learning vision from models rivals learning vision from data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15887–15898. Google Scholar↗

[6] Baydar, M. (2024). *Self-supervised learning for unsupervised image classification and supervised localization tasks* (Doctoral dissertation, Middle East Technical University, Turkey). Google Scholar↗

[7] Xu, C., Wu, J., Zhang, F., Freer, J., Zhang, Z., & Cheng, Y. (2024). A deep image classification model based on prior feature knowledge embedding and application in medical diagnosis. *Scientific Reports, 14*(1), 13244. Google Scholar↗

[8] Gaur, M., Chaturvedi, K., Vishwakarma, D. K., Ramasamy, S., & Prasad, M. (2023). Self-supervised ensembled learning for autism spectrum classification. *Research in Autism Spectrum Disorders, 107*, 102223. Google Scholar↗

[9] Wu, L., Zhuang, J., & Chen, H. (2024). Voco: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22873–22882. Google Scholar↗

[10] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems, 33*, 21271–21284. Google Scholar↗

[11] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical, 17*, 16–23. Google Scholar↗

[12] Alharthi, A. G., & Alzahrani, S. M. (2023). Multi-slice generation sMRI and fMRI for autism spectrum disorder diagnosis using 3D-CNN and vision transformers. *Brain Sciences, 13*(11), 1578. Google Scholar↗

[13] Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering, 30*(4), 2761–2775. Google Scholar↗

[14] Du, Y., Onofrey, J., & Dvornek, N. C. (2024). Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. *arXiv preprint arXiv:2409.18119*. Google Scholar↗

[15] Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., & Bojanowski, P. (2024). DINOv2 meets text: A unified framework for image-and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*. Google Scholar↗

[16] Charoenpanyakul, R., Kittichai, V., Eiamsamang, S., Sriwichai, P., Pinetsuksai, N., Naing, K. M., & Chuwongin, S. (2024). Enhancing mosquito classification through self-supervised learning. *Scientific Reports, 14*(1), 27123. Google Scholar↗

[17] Campanella, G., Chen, S., Verma, R., Zeng, J., Stock, A., Croken, M., Veremis, B., et al. (2024). A clinical benchmark of public self-supervised pathology foundation models. *arXiv preprint arXiv:2407.06508*. Google Scholar↗

[18] Li, S., Cao, Y., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226*. Google Scholar↗

[19] Srivastava, D., Singh, V., Li, S., & Kochersberger, K. (2024). Unmanned Aerial System-Driven Data and Advanced Deep Learning Strategies for Elevating Weed Management in Agricuture. *arXiv preprint arXiv* :23993, 23-45. Google Scholar↗

[20] Morningstar, W., Bijamov, A., Duvarney, C., Friedman, L., Kalibhat, N., Liu, L., & Prakash, S. (2024). Augmentations vs algorithms: What works in self-supervised learning. *arXiv preprint arXiv:2403.05726*. Google Scholar↗

[21] Kumar, P., & Jeyapriyan, M. (2024, September). A novel approach for detection of autism using neural transformer. In *2024 IEEE International Conference on Communication, Computing and Signal Processing (IICCCS)* (pp. 1–6). IEEE. Google Scholar↗

[22] Wang, Y., Pan, K., Shao, Y., Ma, J., & Li, X. (2025). Applying a convolutional vision transformer for emotion recognition in children with autism: Fusion of facial expressions and speech features. *Applied Sciences, 15*(6), 3083. Google Scholar↗

[23] Lewis, R. N., Motiwala, M., Bhura, S., & Mali, S. (2024, November). Comparative study of vision transformers for ASD detection in toddlers using facial features. In *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)* (pp. 500–507). IEEE. Google Scholar↗

[24] Datta, G., Etchart, T., Yadav, V., Hedau, V., Natarajan, P., & Chang, S. F. (2022, May). ASD-Transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4568–4572). IEEE. Google Scholar↗

[25] Rafiki, A., Melinda, M., Oktiana, M., Meutia, E. D., Afnan, A., Mulyadi, M., & Zakaria, L. Q. (2025). Implementation of vision transformer for early detection of autism based on EEG signal heatmap visualization. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, 7*(1), 102–112. Google Scholar↗

*****