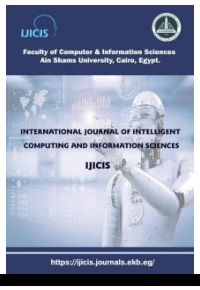
IJICIS, Vol.25, No.2, 54-62
DOI: 10.21608/ijicis.2025.390933.1399

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



A MACHINE LEARNING CASE STUDY ON EARLY DETECTION OF AUTISM SPECTRUM DISORDER USING PHENOTYPIC DATA

Mohamed Gawish*
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mygawish@cis.asu.edu.eg

Nada S. El-Askary
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
nada.sherif@cis.asu.edu.eg

Mohamed Mabrouk Morsey
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mohamed.mabrouk@cis.asu.edu.eg

Abeer M. Mahmoud
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
abeer.mahmoud@cis.asu.edu.eg

Mostafa Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mostafa.aref@cis.asu.edu.eg

Taha Ibrahim El-Arif
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
taha_elarif@cis.asu.edu.eg

Received 2025-05-31; Revised 2025-05-31; Accepted 2025-06-29

Abstract: *Early Autism Spectrum Disorder (ASD) detection is crucial for promoting cognitive, motor skills, and social development. Artificial intelligence-powered systems present an exciting chance to transform ASD detection. The Autism Brain Imaging Data Exchange (ABIDE) represents a significant repository of brain imaging and phenotypic data collected from nineteen sites, encompassing a total of 1,114 cases of both ASD and typical control individuals. Each case includes 347 descriptive variables. This article demonstrates a case study on detecting ASD based on a machine learning (ML) pipeline utilizing phenotypic data from ABIDE. The ML pipeline involves four primary steps: (1) collecting and integrating data, (2) preprocessing the data, (3) training an ML model, and (4) evaluating the ML model. This article employs seven distinct ML algorithms for training the model and documenting the classification accuracy of each algorithm. During the case study, the accuracy ranged from 80.50% to 95.10%. The model trained using the random forest algorithm achieved the preeminent accuracy for ASD detection using phenotypic data.*

Keywords: *ASD, Machine Learning, ABIDE, Phenotypic Data.*

1. Introduction

***Corresponding Author**: Mohamed Gawish
Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt
Email address: mygawish@cis.asu.edu.eg

Mental healthcare improves humans' well-being, functioning, and fulfillment in life [1,2]. Providers, including psychiatrists, psychologists, and other professionals, collaborate to offer personalized support for those facing mental health challenges [3]. One mental disorder that has undergone extensive research is autism spectrum disorder (ASD). It is known as a "spectrum" condition because individuals exhibit a range of symptoms with varying intensity [3,4]. Diagnosing ASD in adults is generally more complex than in children [3]. Moreover, in adults, some symptoms of ASD can resemble those of other mental disorders, such as anxiety disorders and attention-deficit/hyperactivity disorder [3]. As indicated in [5-9], in the United States, one in thirty-six children (8 years old) and one in forty-six children (4 years old) suffered from ASD in 2020.

Figure. 1 demonstrates a marked increase in the prevalence estimates for children aged four and eight years old with ASD per 1000 children in the USA over the twenty years from 2000 to 2020, as reported by the ADDM network [6]. During the twenty years from 2000, the propagation of ASD in children (8 years old) rose by 412%. The weight of early detection of ASD in children is evident, with those aged 4 being 1.3 times more likely than 8-year-olds to be diagnosed. Autism can be identified as early as 24 months [10]. Traditional ASD diagnosis relies on subjective and time-consuming observations by healthcare providers [11].

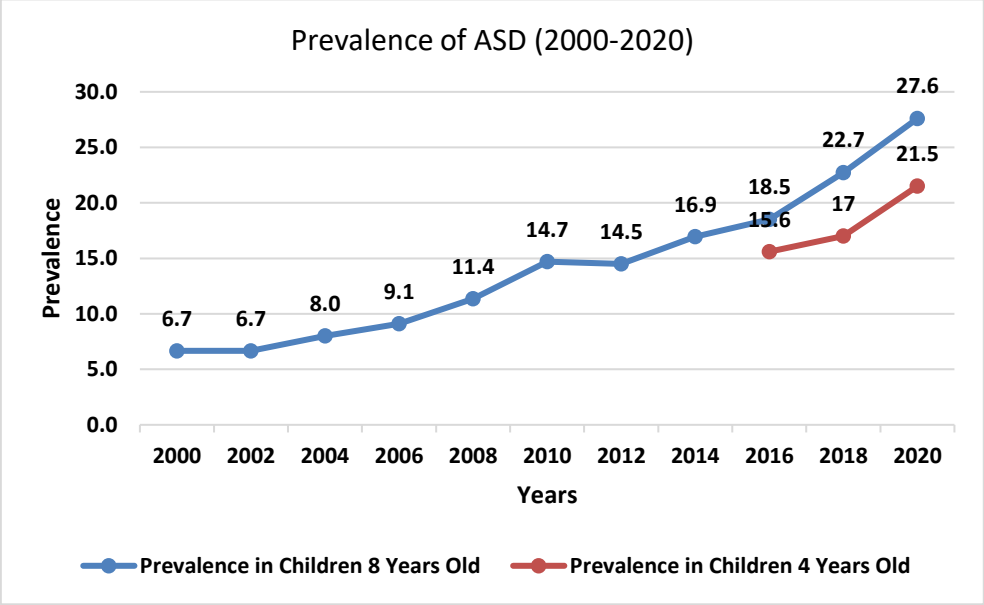


Figure. 1: The spread of ASD over 20 years.

Artificial intelligence (AI) systems have the potential to revolutionize early detection. AI's advanced recognition uses machine learning (ML) algorithms to analyze brain imaging, signal data, paper tests, phenotypic data, behavioral observations, and clinical and genetic data [12]. These algorithms utilize large datasets from both individuals with and without autism, enabling them to effectively identify patterns that may suggest the presence of ASD. Moreover, they leverage natural language processing and computer vision to assess behavioral cues from video and audio recordings of children with ASD.

2. Related Work

Several studies highlight AI's role in early ASD detection and diagnosis, utilizing ML algorithms to: (1) create clinical screening tools for at-risk children, (2) analyze data such as genetic markers, behavioral

observations, and neuroimaging, and (3) improve accuracy in ASD detection. Autism phenotypic data includes observable traits in individuals with ASD, such as communication skills, repetitive behaviors, and sensory sensitivities. It can be collected via observation, caregiver reports, wearable sensors, electronic health records, and standardized assessments. This data helps understand the behaviors associated with ASD and develop targeted interventions. In AI, phenotypic data trains ML algorithms to identify autism patterns. Notable studies and applications using AI for ASD detection using phenotypic data include:

- Tariq et al. [13] introduced a mobile application that uses ML algorithms to analyze short home videos of children. A total of 162 videos (116 ASD, 46 TC) were analyzed, averaging about 2 minutes and 13 seconds in length. ML models were evaluated, including support vector machines (SVM), logistic regression (LR), and decision trees (DT). The best model was LR, achieving 93% accuracy [13]. Validation included independent raters and video sets for unbiased assessment.
- By examining and documenting response-to-name (RTN) behaviors in children, such as response time, consistency, quality, and engagement, Nie et al. [14] implemented ML algorithms to distinguish between the ASD and TC groups based on these characteristics. The research included toddlers diagnosed with ASD (30 cases) and TC (18 cases). The ML model demonstrated consistency rates of 83.33% when compared to ASD diagnoses [14].
- For monitoring emotional states in children with ASD, Talaat et al. [15] introduced a real-time emotion recognition system. The system utilizes a conventional neural network (CNN) for classification in conjunction with a kernel-based sparse autoencoder for feature extraction and dimensionality reduction. It is capable of identifying six facial emotions from images of children's faces taken using smart devices [15]. Additionally, it employs an Internet of Things and fog computing framework to enable swift local processing and alert generation. A dataset consisting of 830 labeled facial images of autistic children was sourced from Kaggle [15]. The study evaluated three pre-trained CNN models: ResNet (91.43%), MobileNet (88.12%), and Xception (95.23%).
- Using eye-tracking data alongside deep learning (DL) models for detecting early-stage ASD by Ahmed et al. [16]. They utilized a publicly accessible dataset containing over 2 million eye-tracking data entries from 59 children (29 ASD and 30 TC). The research involved stringent preprocessing and feature selection methods. Four DL architectures were trained and assessed: Long Short-Term Memory (LSTM), Bi-LSTM, Gated Recurrent Units (GRU), and CNN-LSTM. The models achieved accuracies of 98.33%, 96.44%, 97.49%, and 97.94%, respectively, with the LSTM model showcasing the best performance at 98.33% accuracy [16].
- A comparative study conducted by Elshoky et al. [17] examined how feature selection methods can enhance the classification accuracy of ASD using ML algorithms. The authors analyzed two variations of the AQ-10 dataset sourced from different repositories [17]. Ten distinct classification algorithms were employed to train the ML models: LR, Linear Discriminant Analysis (LDA), Naïve Bayes (NB), SVM, K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Adaboost (AB), GBoost, Random Forest (RF), and Extra Trees (ET). Each model's performance was assessed using ten-fold cross-validation. The leading three classifiers (LR, LDA, and AB) consistently achieved an accuracy of 100% across both dataset versions [17].
- Kunda et al. [18] created a ML model to classify autism based on the ABIDE-I dataset, which includes 1,035 cases (505 ASD, 530 TC) of fMRI and phenotypic data. The model was trained with Ridge, LR, and SVM classifiers, with its effectiveness evaluated through ten-fold and

leave-one-out cross-validation. In the end, the model attained a classification accuracy of 73% using the Ridge classifier [18].

- Gawish et al. [19] explored dimensionality reduction of phenotypic data from the ABIDE-II dataset to enhance the performance of ML classifiers for ASD detection. The dataset includes 1,114 cases, each with 347 behavioral and demographic features. A dual approach to dimensionality reduction was applied: feature selection based on missing data thresholds and principal component analysis (PCA). Three classifiers—DT, SVM, and RF—were trained and evaluated. Feature groups with less than 60% missing data showed the highest classification performance, which achieved a peak accuracy of 94.1% using the RF model. PCA-based models exhibited slightly lower performance, especially when preserving less than 85% of the data variance. The results underscore the importance of strategic feature reduction in improving ASD classification.

The examined studies collectively reveal notable progress in AI-based detection and diagnosis of ASD. They demonstrate how AI can improve the identification and diagnosis of ASD by leveraging ML algorithms. Additionally, these studies highlight certain limitations that could impact the accuracy of ML models in detecting ASD, such as limited sample sizes, imbalanced datasets, poor real-world clinical generalization, and significant computational demands.

3. Methodology

This section describes the ML pipeline and the techniques used for ASD detection. The ML pipeline for ASD detection was presented in [19]. It consists of four sequential steps, as shown in Figure. 2: it starts with data acquisition and integration, then data preprocessing, and ML model training and evaluation.

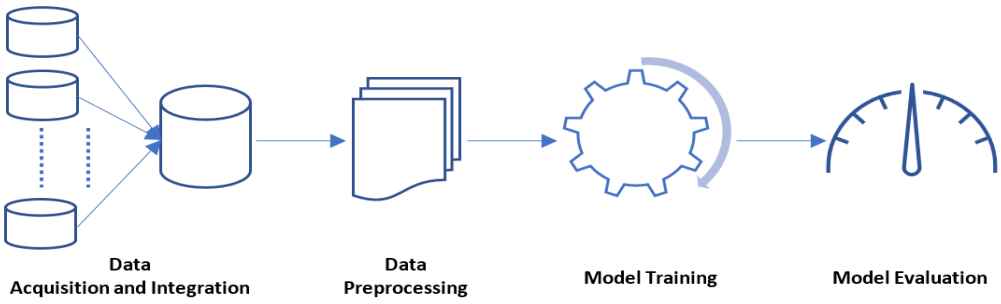


Figure. 2: Machine learning pipeline for detecting ASD [19].

4.1. Dataset

The Autism Brain Imaging Data Exchange (ABIDE) includes neuroimaging and phenotypic data from individuals with and without autism. The publicly available ABIDE dataset [20] has two collections: ABIDE-I and ABIDE-II. Each offers functional and structural MRI images along with phenotypic and anatomical data [20]. ABIDE-I, founded in 2014 across seventeen sites, includes 1,112 cases, ranging between 539 ASD and 573 typical control (TC) cases [21]. ABIDE-I faces challenges such as complex brain connectivity and ASD heterogeneity, emphasizing the need for a larger dataset with distinct characteristics [22].

ABIDE-II, released in June 2016, aims to address these challenges with more phenotypic features and psychiatric variables, especially concerning core ASD and related symptoms [23]. ABIDE-II has 347 features compared to 73 in ABIDE-I. Nineteen international sites contributed 1,114 cases, ranging from 521 ASD to 593 TC cases [23]. The rich detail in ABIDE-II allows researchers to consider not only the autism diagnosis but also symptom variability and co-occurring conditions during imaging analysis.

4.2. Data Integration and Inspection

The model employed phenotypic data from ABIDE-II, which gathered information from over nineteen international sites, making it more comprehensive than any other dataset. Specifically, ABIDE-II includes 1,114 real-world cases of individuals aged 5 to 64 years. The dataset features 521 ASD cases and 593 TC cases. Each case is defined by 347 structured characteristics and clinical data, such as IQ, medical history inquiries, demographic factors, diagnostic status, and a wide range of behavioral assessments. Moreover, essential patient information, including gender, age, handedness index score, and handedness category, is also incorporated.

After downloading the phenotypic data from all nineteen sites involved in the ABIDE-II dataset, a thorough analysis was conducted. The individual datasets were merged into a single file to maintain consistency and uniformity across the various sites. During this analysis, the primary issue identified was the substantial amount of missing data. The dataset comprises 347 features, where the first two columns serve as unique identifiers. The third column indicates the ground truth, assigning a value of 1 for ASD or 2 for TC (later revised to 0; Figure 2 illustrates the changes in labels). This classification results in two categories: ASD with 521 records and TC with 593 records. The remaining 344 columns contain various features reflecting patients' responses to doctors' inquiries. A review of the data revealed several structural issues, prompting the next phase of applying additional data preprocessing techniques.

4.3. Data Preprocessing

Essential preprocessing techniques were employed to enhance the quality of subsequent analyses while maintaining and improving data balance. Data cleaning must be executed meticulously to avoid ending up with either insufficient or excessive irrelevant data. This procedure was conducted in several stages:

- Duplicate observations were removed.
- Missing data was addressed by either filling in suitable values or excluding entire observations, based on the features of the missing data and the model's requirements.
- Undesirable outliers were filtered out. Some values, which ranged from -100 to 100, were adjusted by adding 100 to convert them into a non-negative range of 0 to 200.
- Non-numeric values were manipulated. The raw dataset includes six columns with string data types, all of which needed conversion into numeric values to align with the model's requirements. The Levenshtein distance algorithm was applied for this conversion.

4.4. Model Training and Evaluation

The training data input for the model consists of a feature matrix that contains instances representative of individuals with ASD and typical control subjects, with each instance defined by carefully chosen features. Following this, a testing phase is executed using a dataset that the model has not seen before to measure its performance. Ultimately, when faced with new instances, the model should recognize individuals with ASD based on their specific features.

4. Case Study

This section performs a case study on utilizing ML for the detection of ASD through the ABIDE-II dataset. This case study aims to assess how effectively ML classifiers differentiate between individuals with ASD and those who are typically developing. The case study is structured as follows:

5.1. Dataset

This study utilized a subset of the ABIDE-II phenotypic dataset. We randomly selected 50% of the available cases, yielding 557 participants: 261 ASD individuals and 296 with typical development. The data was then split into an 80/20 division for training and testing sets. To maintain a balanced analysis, the training set included 446 participants, whereas the testing set had 111 participants.

5.2. Tools and Implementation

The classification task employed various ML classifiers, all executed within the MATLAB programming environment. Each classifier was set up using default parameters to evaluate performance. Detailed configurations for each model are provided below:

- Random Forest (RF):
  The RF classifier comprised thirty decision trees. At each decision node, a random subset of predictors was chosen. Sampling the square root of the number of observations as predictors at each node. The maximum number of splits allowed per tree matched the total number of observations, enabling fully grown trees. These parameters balance model complexity and efficiency while ensuring strong performance in training and testing.

- Support Vector Machine (SVM):
  The separating hyperplane is created by a linear kernel function. This design was selected due to its simplicity and efficiency in handling high-dimensional spaces.

- Fine K-Nearest Neighbors (Fine KNN):
  This model utilized a single nearest neighbor (k=1) along with the Euclidean distance metric.

- Decision Tree (DT):
  The decision tree classifier allowed a maximum of 100 splits, using Gini's diversity index as the split criterion for node partition quality evaluation.

- Fine Gaussian Support Vector Machine (Fine Gaussian SVM):
  A Gaussian kernel (radial basis function) was employed to enable nonlinear decision boundaries within the feature space.

- Weighted K-Nearest Neighbors (Weighted KNN):
  This version of the KNN algorithm utilized ten nearest neighbors (k=10) along with the Euclidean distance metric. The distance weights were calculated using a squared inverse function, which emphasized the impact of closer neighbors during analysis classification.

- Random Forest with Principal Component Analysis (RF + PCA):

Feature reduction was conducted through PCA before training the RF model. PCA components were chosen to maintain 95% of the total dataset variance. The RF classifier was set up the same way as the standard RF model.

Seven classifiers were trained individually and then assessed with the testing dataset. The accuracies recorded were 95.10% for RF, 93.30% for SVM, 87.30% for KNN, 94.00% for DT, 80.50% for fine Gaussian SVM, 83.30% for weighted KNN, and 86.20% for RF+PCA, as depicted in Figure. 3. The RF model provided the highest classification accuracy. This case study highlights the crucial role of data preprocessing.

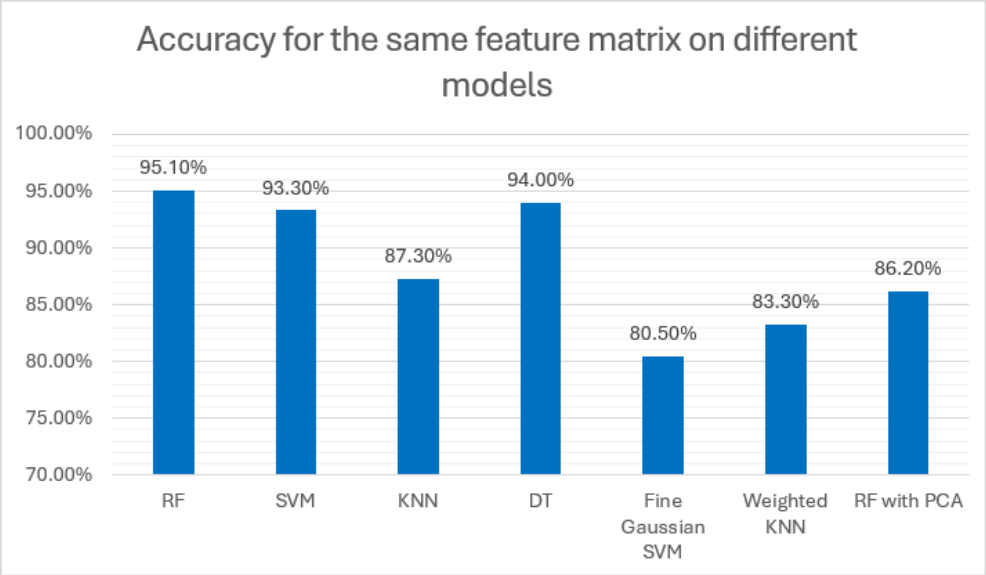


Figure. 3: Accuracy of the different ML models involved in the case study.

5. Conclusion and Future Works

The studies reviewed collectively reveal substantial progress in AI-assisted ASD diagnosis while also pointing out major limitations related to dataset availability, model generalization, and computational feasibility. The case study showed that the RF model attained the top accuracy of 95.10% among all models. This result corroborates existing literature, which frequently notes the superior performance of ensemble methods in high-dimensional healthcare datasets.

The significant findings of this study not only validate the efficacy of ML, particularly RF, in ASD detection but also emphasize the crucial role of data preprocessing in achieving optimal classification outcomes. This aligns with past research that indicates how inadequate data preparation can severely impair model performance, irrespective of the classifier utilized. The model's success relies strongly on the quality and availability of phenotypic data. These insights provide a beneficial roadmap for future research and potential clinical uses of AI-driven ASD screening tools.

References

[1] V. Singh, A. Kumar, S. Gupta, "Mental Health Prevention and Promotion-A Narrative Review," Frontiers in Psychiatry, 13 (2022) 898009. https://doi:10.3389/fpsyt.2022.898009

[2] S. Saxena, P.K. Maulik, "Prevention and Promotion in Mental Health," Geneva, Geneva: World Health Organization, 2002.

[3] Autism Spectrum Disorder, National Institute of Mental Health. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd (accessed 6 March 2025)

[4] Autism Spectrum Disorder, Mayoclinic.org – Diseases & Conditions, https://www.mayoclinic.org/diseases-conditions/autism-spectrum-disorder/symptoms-causes/syc-20352928 (accessed 6 March 2025)

[5] Data & Statistics on Autism Spectrum Disorder, Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/ncbddd/autism/data.html (accessed 11 March 2025)

[6] M. J. Maenner, Z. Warren, A. R. Williams, et al., "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020," MMWR Surveillance Summary, 72(No. SS-2) (2023) 1–14.

[7] K. A. Shaw, M. J. Maenner, J. Baio, et al., "Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years — Early Autism and Developmental Disabilities Monitoring Network, Six Sites, United States, 2016." MMWR Surveillance Summary, 69(No. SS-3) (2020) 1–11.

[8] K. A. Shaw, M. J. Maenner, A. V. Bakian, et al., "Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018." MMWR Surveillance Summary, 70(No. SS-10) (2021) 1–14.

[9] K. A. Shaw, D. A. Bilder, D. McArthur, et al., "Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020." MMWR Surveillance Summary, 72(No. SS-1) (2023) 1-15.

[10] C. Okoye, C. M. Obialo-Ibeawuchi, O. A. Obajeun, S. Sarwar, C. Tawfik, et al., "Early Diagnosis of Autism Spectrum Disorder: A Review and Analysis of the Risks and Benefits," Cureus, 15(8) (2023).

[11] I. Jamwal, D. Malhotra, M. Mengi, "Autism Spectrum Disorder Detection Using ASD_sfMRI," In: Computer Vision and Robotics: Proceedings of CVR 2021, Springer, Singapore, 2022, p.175-189.

[12] A. A. Lima, M. F. Mridha, S. C. Das, M. M. Kabir, M. R. Islam, Y. Watanobe, "A Comprehensive Survey on the Detection, Classification, and Challenges of Neurological Disorders," Biology, 11(3), 2022, p.469.

[13] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, D. P. Wall, "Mobile detection of autism through machine learning on home video: A development and prospective validation study," PLOS Medicine, 15(11) (2018) e1002705.

[14] W. Nie, B. Zhou, Z. Wang, B. Chen, X. Wang, C. Hu, H. Li, Q. Xu, X. Xu, H. Liu, "Computational interpersonal communication model for screening autistic toddlers: A case study of response-to-name," IEEE Journal of Biomedical and Health Informatics, 28(6) (2024) 3683-3694.

[15] F. M. Talaat, Z. H. Ali, R. R. Mostafa, N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," Soft Computing, 28(9) (2024) 6695-6708.

[16] Z. A. Ahmed, E. Albalawi, T. H. Aldhyani, M. E. Jadhav, P. Janrao, M. R. M. Obeidat, "Applying eye tracking with deep learning techniques for early-stage detection of autism spectrum disorders," Data, MDPI, 8(11) (2023) 168.

[17] B. Elshoky, O. Ibrahim, A. Ali, "Machine Learning Techniques Based on Feature Selection for Improving Autism Disease Classification," International Journal of Intelligent Computing and Information Sciences, 21(2) (2021) 65–81.

[18] M. Kunda, S. Zhou, G. Gong, H. Lu, "Improving Multi-Site Autism Classification via Site-Dependence Minimization and Second-Order Functional Connectivity," IEEE Transactions on Medical Imaging, 42(1) (2022) 55‑65.

[19] M. Gawish, N. S. El-Askary, M. M. Morsey, A. M. Mahmoud, M. Aref, T. I. El-Arif, "Dimension Reduction of Phenotypic Data for Enhancing Autism Spectrum Disorder Detection," In: the 2nd IEEE International Conference on Machine Intelligence and Smart Innovation (ICMISI 2025), 2025.

[20] Autism Brain Imaging Data Exchange. https://fcon_1000.projects.nitrc.org/indi/abide/ (accessed 28 December 2024)

[21] A. Di Martino, CG. Yan, Q. Li, et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," Molecular Psychiatry, 19(6) (2014) 659–667 https://doi.org/10.1038/mp.2013.78

[22] C. M. Williams, H. Peyre, R. Toro, A. Beggiato, F. Ramus, "Adjusting for allometric scaling in ABIDE I challenges subcortical volume differences in autism spectrum disorder," Human Brain Mapping, 41(16) (2020) 4610‑4629. https://doi.org/10.1002/hbm.25145

[23] A. Di Martino, D. O'Connor, B. Chen, K. et al., "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," Scientific Data, 4(1) (2017) 1-15. https://doi.org/10.1038/sdata.2017.10