Prediction of Ethnic Influences on Autism Spectrum Disorder: A Multifaceted Integrative Analysis of Clustering and Artificial Neural Networks

Dhruv Sharma
22228268
MSc Data Analytics
National College of Ireland

Abstract

This research project aims to examine the discrepancy in proper classification models for identification of autism spectrum disorder (ASD) in participants. Through thorough application of Artificial Neural Networks (ANN) and clustering analysis based on participants' country of residence as part of the methodology, this project explores its effectiveness compared to other classification models. Additionally, this project ventures to employ the proposed model to facilitate individuals with ASD through a digital user interface, recommending personalized intervention strategies for enhancing their social skills and overall standard of living. This project assimilates knowledge from psychology, data science, and social science to impart valuable knowledge into ASD diagnosis, intervention methods, and individualized techniques. The proposed ANN model will furnish a more comprehensive understanding of effective intervention methods and clustering analysis will present clarity upon the ethnic influences on ASD which would deliver beneficial information for researchers, psychologists, and guardians having children diagnosed with ASD.

Keywords: Artificial Neural Networks, Clustering Analysis, ASD, Ethnicity

1. Introduction

Autism Spectrum Disorder is a neurodivergent disease known to showcase substantial hurdles in diagnosis and early intervention because to its intricate characteristics. Despite enormous research investigations, a notable void remains in the availability of appropriate classification models for early identification of ASD amongst individuals. However, traditional methods of ASD detection can be time-consuming and often lack the accuracy needed for early diagnosis. This project aims to fill this void by employing both supervised and unsupervised machine learning algorithms such as artificial neural networks as a classification model and, clustering based solutions by comparing its effectiveness with alternate classification methods for more effective identification. Additionally, this project reveals practical observations into early identification and intervention strategies for enhancing social skills and overall standard of living for people diagnosed with ASD.

The primary research questions revolve around evaluating the effectiveness of clustering analysis based on participants' country of residence and ANN models in predicting ASD scores to determine autism. The research problem at present revolves around the impediments of current ASD diagnostic methods and the

potential of machine learning application to address these shortcomings. Conventional approaches mostly rely on subjective assessments and empirical data, leading to variability in diagnosis and treatment result. Contrastingly, machine learning methods offer a data-driven approach to analysing patterns and identifying predictive features, which could crucially enhance the accuracy and efficiency of ASD diagnosis and personalised intervention strategies could be handed out to families of subjects using a recommender system which feeds on data at hand. Synergizing clustering analysis into the process allows for the recognition of definite subgroups within ASD populations based on ethnicity and other features, providing valuable insights for personalized intervention planning.

This study's objective to give a dissertation to fundamental questions concerning the productiveness of machine learning techniques in improving autism spectrum disorder detection and the potential synergistic advantages of putting together clustering analysis and Artificial Neural Networks (ANN) for a higher success rate as well as unveiling if demographic factors have influence over ASD. At its core, the research endeavours to investigate how clustering analysis, particularly based on participants' country of residence, can reveal nuanced subtypes within the ASD population. Subsequently, the study aims to discover how ANN models can harness the data at hand to enhance diagnostic accuracy and facilitate accurate predictions of positive and negative ASD subjects. With the help of this analysis, the study aims to come up with some advance protocol for screening ASD subjects more efficiently by combining multiple machine learning disciplines together. Also, the development of recommender system which caters to the individualized needs of ASD patients across different geographical regions can turn out to be a significant aid for the subjects. Persuading the use of ML techniques for ASD detection stems out of their capacity to leverage large datasets and identifying complex patterns in it to correctly make predictions. ANN models have shown success in multiple domains, including healthcare, by constructively learning from data and making information driven decisions. By channelling the prowess of clustering analysis and ANN, researchers can develop sophisticated ASD identification systems which keeps individual differences in consideration while making predictions. This study focuses on investigating how effective ML techniques are in detecting ASD and incorporating clustering analysis and ANN into current diagnostic protocols. The study's specific goal is to investigate how clustering analysis can pinpoint different ASD subtypes and how ANN models can utilize this information to enhance diagnostic accuracy. Moreover, the study aims to explore the possible advantages of combining clustering analysis and ANN in the field of clinical practice, specifically in relation to personalized intervention planning and predicting outcomes.

In addition to employing Artificial Neural Networks (ANN), this research project incorporates clustering analysis and thorough Exploratory Data Analysis (EDA) to enhance the accuracy and effectiveness of ASD detection and intervention. Clustering analysis enables organizing participants based on their country of residence which aids researchers in identifying potential regional differences in ASD prevalence and symptomatology. By analysing clusters, researchers can gain valuable insights into geographic variations in ASD and tailor personalised intervention strategies accordingly. Moreover, EDA plays a crucial role in understanding the underlying patterns and relationships within the dataset, identifying relevant features, and preprocessing data for ANN modelling. Through comprehensive EDA techniques such as data visualization, correlation analysis, and feature engineering, researchers can optimize the input variables for ANN models, thereby improving their predictive performance. Integrating clustering analysis and EDA into the research methodology enhances the robustness and interpretability of the ANN-based ASD detection system,

providing clinicians and educators with actionable insights for early intervention and personalized treatment planning.

The possible impacts of this work are noteworthy. The study seeks to improve ASD research by creating more precise and individualized diagnostic tools using clustering analysis and ANN models. Moreover, incorporating ML methods in clinical settings may transform ASD treatment approaches, resulting in improved results for individuals with ASD and their loved ones. In the end, the suggested solution could potentially close the divide between research and practice by presenting original methods for detecting and intervening in ASD that prioritize the unique needs and traits of everyone. In terms of organization, this document is structured to offer a thorough examination of the research project, emphasizing the importance of clustering analysis and ANN models in detecting and intervening in ASD. Every part adds to the one before it, resulting in a unified story detailing the research method and its anticipated impact. By incorporating background information, research goals, methodology, and potential advantages, this paper seeks to offer a convincing explanation for the planned study.

2. Literature Review

The literature review section of this analysis inspects prior research on the utilization of machine learning techniques employed in autism spectrum disorder early identification and diagnosis. By utilizing existing literature, the objective of this project is to identify void, hurdles, and opportunities for extended research in the domain of ASD diagnosis and management. The review will differentiate key studies to elucidate the strengths and limitations of different ML approaches, finally setting the stage for the proposed research.

A. Past works utilizing ML procedures for ASD detection.

In earlier studies, numerous ML algorithms have been employed for ASD screening, each contributing unique insights and methodologies. For instance, Hemu et al. (2022) utilized algorithms including Random Forest and XGBoost to predict ASD based on extensive datasets, achieving great accuracy rates. The authors procedure relied on feature identification and ensemble methods to improve the model's evaluation on unseen data. In contrast, Safahi et al. (2024) highlighted the significance of feature engineering in ML-based ASD screening, stressing the role of domain-specific knowledge in selecting relevant features. Both studies exhibited the potential of machine learning in autism identification, however, the authors varied in their approaches to feature selection and model optimization.

B. Challenges and Limitations of Traditional Diagnostic Methods.

Conventional diagnostic methods for ASD face many challenges, including subjectivity and time-consumption. Karim et al. (2021) highlighted the limitations of behavioural observation and clinical assessment, emphasizing the need for more objective and efficient diagnostic tools. Contrarily, Shetty et al.

(2022) discussed about the reliability and accountability of conventional approaches, warning against the dependence on ML algorithms, suggesting not to rely solely on them without considering clinical inference. By comparing these viewpoints, it becomes clear that while machine learning methods enhance diagnostic accuracy, traditional screening techniques supply valuable insights on an individual level.

C. Integration of ML in Healthcare Practices.

The incorporation of ML techniques into healthcare sector presents both opportunities and hurdles in ASD diagnosis and management. Almana and Hammad (2022) inspected the possible benefits of ML-based screening methods in enhancing diagnostic accuracy and simplifying clinical screening workflows. The author's research pointed out the adaptability and effectiveness of ML algorithms in processing large datasets and recognize subtle patterns indicative of ASD. However, Farooqui et al. (2022) raised concerns about the ethical implications of ML integration, specifically with information security and model interpretability. Despite these hindrances, the combination of machine learning in healthcare holds promise for improving ASD diagnosis and intervention strategies, suggesting the need for further research in this area.

D. Progressions in machine learning techniques for ASD diagnostics.

Recent progressions in machine learning have notably contributed to the domain of autism spectrum disorder (ASD) diagnostics, supplying novel insights and methodologies. For example, Kamala et al. (2021) suggested a ML-based strategy for ASD identification, making use of various machine learning techniques including feature selection techniques. Their study revealed favorable results in correctly identifying ASD subjects across different age groups, and placed emphasis on the capabilities of ML in enhancing diagnostic accuracy. Moreover, Ningsih et al. (2021) developed an Android-based app, ASD Detector, utilizing CHAT analysis and Certainty Factor method for diagnosing ASD in children. This innovative application confirmed great accuracy rates in providing diagnostic results compatible with expert assessments, emphasizing the need for inclusion of data driven approaches in ASD detection. Furthermore, Akter et al. (2021) introduced a machine learning-based classification model for early detection of autism, concentrating on cleansing feature sets and leveraging them for the model. Their strategy yielded greater performance in predicting ASD, highlighting the significance of continuous refinement and optimization in ML-based diagnostic tools. These studies conjointly showcase the swift progress and assorted applications of ML methods in advancing ASD diagnosis, making the way for more productive and accessible diagnostic solutions.

E. Identification of Research Niche and Expected Contribution.

In outline, the reviewed literature supplies valuable insights into the current landscape of ML-based ASD detection and intervention techniques. However, there exists a research niche in the integration of ML algorithms like artificial neural networks and clustering for outcome prediction as well as planning intervention strategies on an individual scale. By addressing this void, the proposed research's objective is to provide to the development of better and increasingly effective diagnostic tools for ASD detection, ultimately decreasing the false positive rates in ASD screening making the process as accurate as possible. Recent advancements in ASD identification have seen the integration of unconventional technologies and machine learning methodologies to improve diagnostic precision. Simon et al. (2024) presented a ground-breaking approach that consolidates machine learning and deep learning techniques with gaze tracking for pre-emptive identification of ASD, utilizing visual hints caught through gaze tracking technology. Kulyabin et al. (2024) proposed a gated multilayer perceptron (MLP) architecture for ASD classification using electroretinogram (ERG) data, with fusion of attention mechanisms to concentrate on appropriate features extricated from ERG waveforms. Additionally, Mittal et al. (2024) investigated decision tree classification methods for autism risk evaluation, expanding predictive models using machine learning algorithms to evaluate the plausibility of ASD based on socio-demographic and behavioural elements. These innovative methodologies underline the prowess of leveraging advanced technologies and computational methods to improve ASD detection and risk assessment, hence smoothening early intervention and support strategies.

3. Research Method & Specification

The proposed solution to the research question incorporates the application of both supervised and unsupervised algorithms for early identification and future detection of autism spectrum disorder. Particularly, a multifaceted integrative approach will be employed by integrating knowledge from data science, psychology, and neuroscience. This project strategy involves exploratory data analysis of ASD-related dataset, attribute selection, model training and testing using ML algorithms and final evaluation of the proposed model. Clustering will also be incorporated as a secondary model to identify whether an individual's ethnicity depicts any significant part in ASD subjects. This impactful observation will assist to discover the influence of demographic factors over the probability of a person diagnosed with autism spectrum disorder. Also, the tools identified for use in this project research includes Python programming language for libraries such as TensorFlow, Scikit-learn, and Keras for ML model development and evaluation. For larger dataset, Apache 2.0 will be utilised. Weka software will be employed for data mining activities for further research.

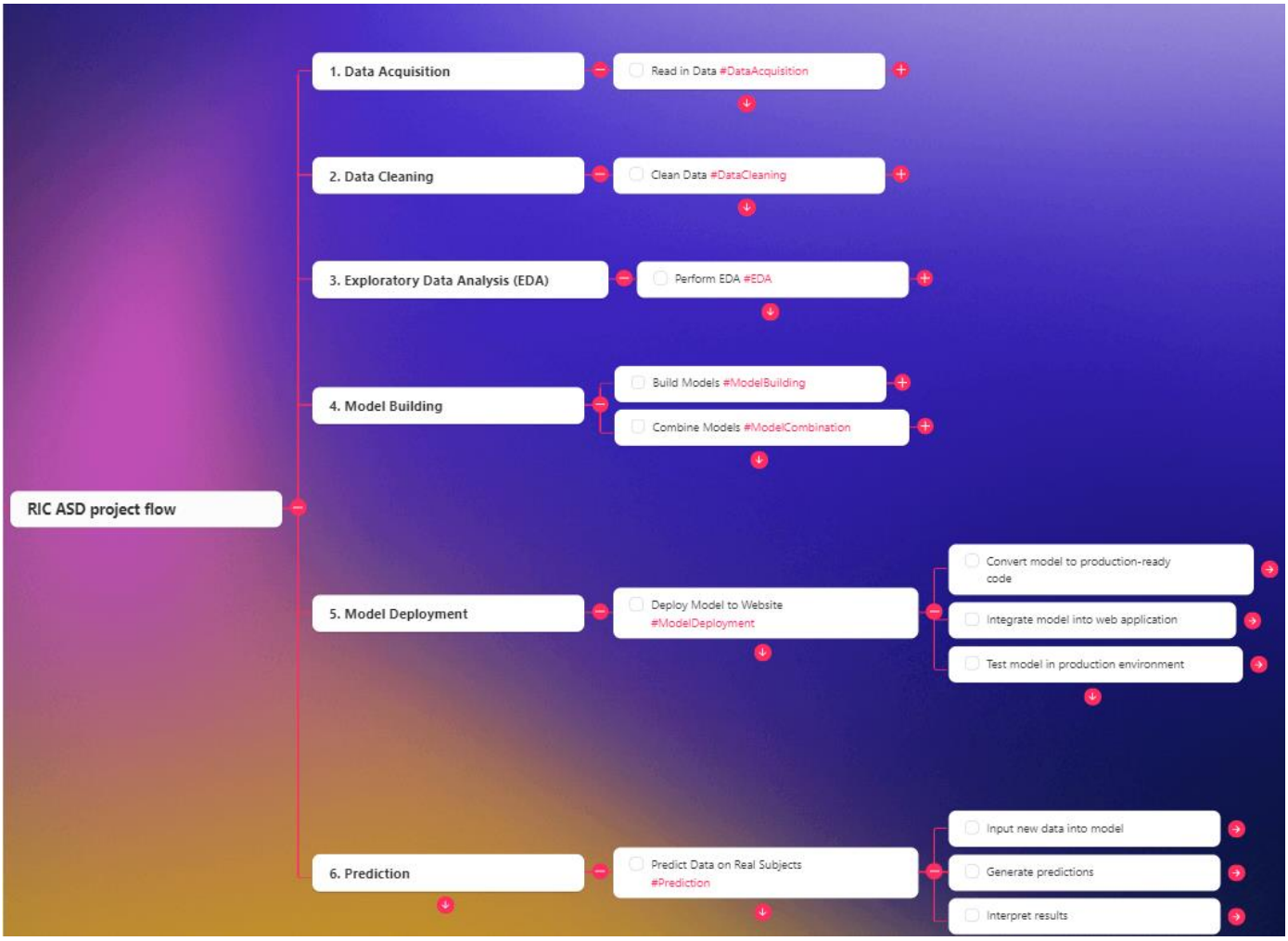

Figure 1 : Project Flow Diagram

A. Dataset Description and Preprocessing.

The dataset comprises demographic information of 705 participants of autism screening conducted by the national institute of health research. This dataset comprises of 20 attributes including participant's country of residence, age, gender, ethnicity, gender, and ASD result scores. It encompasses records from a diverse stratum, with data variables representing both categorical and numerical features creating a multivariate dataset. Earlier research studies conducted with this dataset showcase improper data handling which resulted in an unfit classification model. The preprocessing step involves removing outliers, handling missing values to balance the dataset, encoding categorical variables for instance using one-hot encoding on 'country of residence'), and scaling numerical columns to secure model unbiasedness due through normalisation. Furthermore, a correlation matrix was generated for essential feature selection before fitting the proposed model.

B. Clustering Algorithm Selection

Various categories of clustering algorithms will be deployed for effective identification of whether a demographic of participants possess any influence over the condition. Hierarchical clustering will be utilized for its capability to handle non-linear corelation to generate hierarchical clusters in the data. This method selection is justified because of the heterogeneous nature of the dataset and the requirement to identify distinct subgroups based on the participant's geographic location.

C. ANN Architecture and Training

The ANN modelling will incorporate multiple number of layers, encompassing the input layer, hidden layer, and output layer. The count of perceptions in each subsequent layer, activation functions such as ReLU and sigmoid functions, and regularization techniques like dropout will be observed based on the proposed model's performance. Training the model will be fulfilled through the backpropagation technique with stochastic gradient descent optimization by investigating methods such as learning rate scheduling and early stopping to hinder model overfitting.

D. Evaluation and Deployment Strategy

The quality of clusters will be evaluated by employing the silhouette score result and through visual demonstration of dendrograms. The ANN model's efficacy will be evaluated by using metriculation such as model accuracy, precision score, recall count, F1-score result, and by analysing the ROC curve. The trained ANN model will be fine-tuned by utilising cross validation method to assure better results on unseen data which will be deployed in a digital web-based user interface. This provides for a scalable and real-world prediction model for identification of autism in new participants. Some regular considerations will involve the model's computational performance and the need for regular weekly updates. Statistical methods will be used to carefully analyse and evaluate the results of the experimental research, offering proof to either confirm or refute the hypothesis.

E. Ethical Considerations

The ethical concerns around this research project includes participant's data privacy, their informed consent to use their information, and the responsible application of machine learning in the domain of mental health. The data comprises of 705 participants and was made available by the National Institute for Health Regulations.

F. Integration of Clustering and ANN Model

The results of clustering analysis combined with the features set used in fitting and training for the ANN model will be combined as an additional input feature which represents the new cluster. This approach will improve the model's capability to discover unidentified patterns. Another methodology is ensemble learning, where the results from both clustering analysis and the ANN model are combined to improve overall model performance.

G. Project Plan

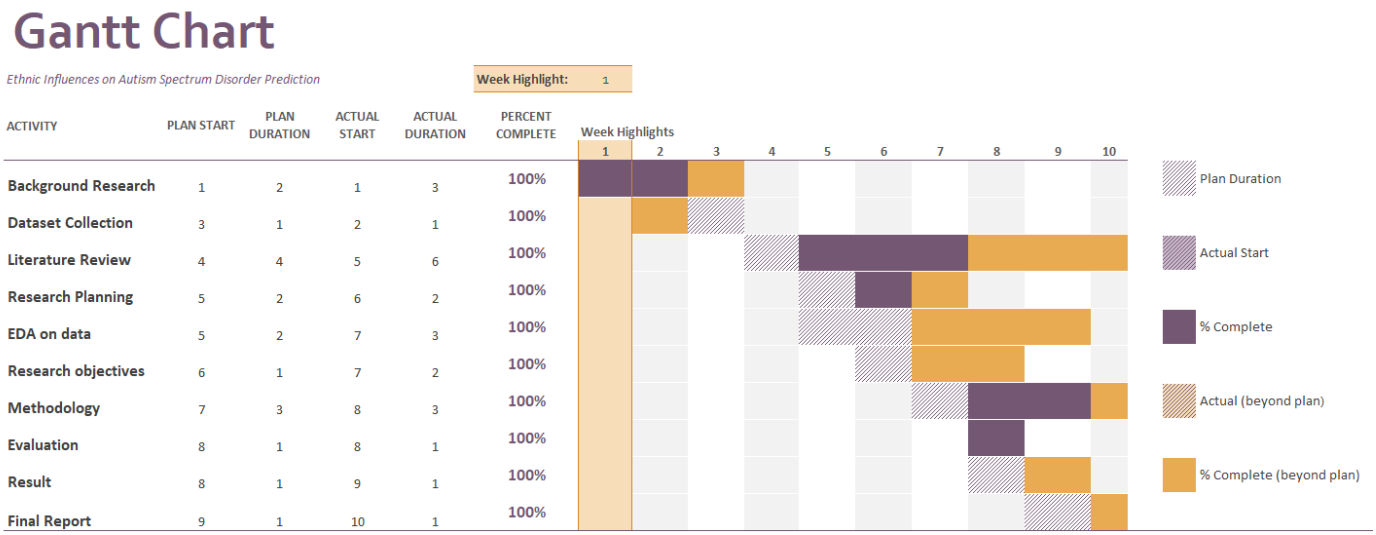

Figure 2 Gantt Chart

Lastly, this research proposal presents a new method for diagnosing ASD by combining clustering analysis and ANN modelling. By conducting thorough testing and evaluation, the main objective is to show how effective this approach is in enhancing diagnostic accuracy and guiding personalized intervention strategies for positive subjects.

REFERENCES

R. B. Hemu, M. M. Mim, M. M. Ali, K. Nayer, K. Ahmed, and F. M. Bui, "Identification of Significant Risk Factors and Impact for ASD Prediction among Children Using Machine Learning Approach," in 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2022, pp. 1-6.

S. Almana and M. Hammad, "Predicting ASD Using Optimized Machine Learning," in 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 2022, pp. 1598-1602.

B. Kamala, K. S. Mahanaga Pooja, S. Varsha, and K. Sivapriya, "ML Based Approach to Detect Autism Spectrum Disorder (ASD)," in 2021 4th International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2021, pp. 313-318.

S. Karim, N. Akter, M. J. A. Patwary, and M. R. Islam, "A Review on Predicting Autism Spectrum Disorder (ASD) meltdown using Machine Learning Algorithms," in 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 2021, pp. 1-6.

S. Ningsih, I. Tahyudin, and S. A. Sholikhatin, "ASD Detector: Android-based App Innovation to Detect Autism Spectrum Disorder on Children using CHAT Analysis and Certainty Factor," in 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), Makasar, Indonesia, 2021, pp. 1-7.

Z. Safahi, E. Azimipour, S. Saedi, and S. Sulaimany, "Improving Machine Learning based ASD Diagnosis with Effective Feature Selection," in 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), Babol, Iran, Islamic Republic of, 2024, pp. 1-6.

T. Shetty, V. Zope, M. Dandekar, A. Devnani, and P. Meghrajani, "sha-Early Intervention for children at risk of ASD," in 2022 International Conference on Industry 4.0 Technology (I4Tech), Pune, India, 2022, pp. 1-5.

Q. A. Farooqui, M. A. Rahman, and Shabbar, "Autism Spectrum Disorder (ASD) Diagnosis and Reinforcement by Machine Learning and Neural Networks," in 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 444-451.

V. Vishal, A. Singh, Y. B. Jinila, K. C, S. P. Shyry, and J. Jabez, "A Comparative Analysis of Prediction of Autism Spectrum Disorder (ASD) using Machine Learning," in 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1355-1358.

T. Akter, M. I. Khan, M. H. Ali, M. S. Satu, M. J. Uddin, and M. A. Moni, "Improved Machine Learning based Classification Model for Early Autism Detection," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh, 2021, pp. 742-747.

Y. Srikar, V. Himaanshu, G. Suryanarayana, K. LNC Prakash, and B. V. Saketha Rama, "A Comparative Analysis: Autism Spectrum Disorder Among Children Using Classification Models," in 2023 2nd International Conference on Ambient Intelligence in Health Care (ICAIHC), Bhubaneswar, India, 2023, pp. 01-06.

J. Simon, N. Kapileswar, D. M, M. S., & K. D. G., "Gaze-Assisted Autism Spectrum Disorder Identification: A Fusion of Machine Learning and Deep Learning Approaches for Preemptive Identification," in 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), Lalitpur, Nepal, 2024, pp. 525-529. doi: 10.1109/ICMCSI61536.2024.00082.

M. Kulyabin, P. A. Constable, A. Zhdanov, I. O. Lee, D. A. Thompson, & A. Maier, "Attention to the Electroretinogram: Gated Multilayer Perceptron for ASD Classification," in IEEE Access, vol. 12, pp. 52352-52362, 2024. doi: 10.1109/ACCESS.2024.3386638.

K. Mittal, K. S. Gill, D. Upadhyay, V. Singh, & S. Aluvala, "Applying Machine Learning for Autism Risk Evaluation Using a Decision Tree Classification Technique," in 2024 2nd International Conference on Computer, Communication and Control (IC4), Indore, India, 2024, pp. 1-6. doi: 10.1109/IC457434.2024.10486622.