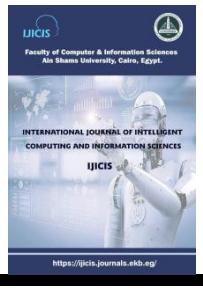
IJICIS, Vol.25, No.2, 32-41
DOI: 10.21608/ijicis.2025.378373.1390

International Journal of Intelligent Computing and Information Sciences

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



PREDICTING EARLY-STAGE AUTISM SPECTRUM DISORDER USING SYMPTOMATIC DATASET FOR ARAB CHILDREN

Dina Ayman Abu Taleb*

Computer Science Department,
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
Dina.ayman@cis.asu.edu.eg

Mohmed Mabrouk Morsey

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

Manal Omar

Medical Studies Department,
Faculty of Postgraduate Childhood Studies, Ain Shams University
Cairo, Egypt
manalomar@chi.asu.edu.eg

El-Sayed M. El-Horbaty

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

Received 2025-04-23; Revised 2025-06-13; Accepted 2025-06-30

Abstract: *The early diagnosis process of autism spectrum disorder (ASD) in toddlers is critical and needs high experience and time to ensure that the diagnosis is accurate and the child has ASD. The early diagnose of ASD can help limit the development of the condition and provide a better life to the patients. To achieve this, many researchers studied how to apply machine learning (ML) algorithms in developing prediction models that help in early diagnosis of ASD. In this research, we leveraged our collected dataset that focuss on Arab children who have ASD especially in Egypt to develop prediction model using ML algorithms, comparing two data splitting approach: 10-fold cross-validation and train-test split. We evaluated the accuracy of Naïve Bayes, Decision trees, Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN). Experimental results show cross-validation achieved an accuracy of 94.92% for Naïve bayes, 87.81 for decision trees, 94.41% for SVM, 92.38% for LR, and 96.44% for ANN algorithm, while train-test achieved 93.22% for Naïve Bayes, 88.13% for decision trees, 91.52% for SVM, 89.83% for LR, and 91.52% for ANN.*

Keywords: *Autism Spectrum disorder (ASD), Machine learning, algorithms, Symptomatic dataset, k-fold cross validation*

***Corresponding Author**: Dina Ayman Abu Taleb

Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: Dina.ayman@cis.asu.edu.eg

1. Introduction:

In recent years, reported numbers for autism spectrum disorder (ASD) roughly 1 in 100 children across the globe. This percentage is influenced by some factors such as the geographic distribution, the number of children in the country, diagnostic services, and diagnostic practices [1].

Autism spectrum disorder is defined as a neurodevelopmental disorder in the developmental period which is associated with other disorders. For example, intellectual developmental disorder, attention deficit/ hyperactivity disorder (ADHD), learning disorder, catatonia, Rett syndrome, environmental factors, medical or genetic conditions, anxiety, depressive disorders, bipolar disorders, or sleeping disorders [2]. The main features that lead to diagnosing the patient as autistic, social communication and interaction are persistent impairment, repetitive behavior, interests, or activities, and limited daily functioning [2].

These features start to be evident in early childhood during the age of 12–24 months or maybe before 12 months. In fact, ASD is not a degenerative disorder, so patients can continue their lives and some of them can work and live independently in adulthood [3]. This is achieved by developing some aspects of the patients, e.g. social communication.

In some cases, with a high severity level of ASD, it is hard to live independently without help from family members even in adulthood. The diagnosing process of ASD is laborious, inefficient, and requires a lot of time [4], in which the expert observes the child's relationship using structured and unstructured activities with their parents and unknown individuals for the child [5]. This process can be done through several tools that are used to diagnose ASD at an early stage, but the time required to get the result is too long. For example, the Screening tool for Autism in Toddlers and Young Children (STAT) requires 20 minutes and the Autism Diagnostic Observation Schedule (ADOS) requires 45 minutes, which is too long a time [5]. Due to the long period of time required to diagnose a child, which is boring and exhausting for children, parents, and doctors. This can affect the daily life of the child and the daily treatment process in which the child needs to take some medicine and have a therapy session that improves the social communication, behavior and language speech of the child.

Under those circumstances, the speed-up of the diagnosis process is important for the patient with ASD, so using machine learning (ML) to build a tool that can be used to diagnose ASD without consuming time and with high accuracy is an imperious need nowadays.

There is a lot of research conducted to reduce the disadvantages of old diagnosis methods such as time consuming and limited accessibility, and tried to use ML algorithms to build an accurate and efficient tool [4,6] which can improve the detection, diagnosis, and prediction process and support the clinicians work and make it fast.

In this study, we tried to address the use of ML algorithms to build a model that can predict ASD in children at an early age, using a new dataset collected from different places that provide care for autistics. The contributions we provided in this study are as follows:

- We have built a new dataset for toddlers with ASD from different centers that provide care for autistic children.
- We have flittered the features of the dataset and choose the best ones that helped us in building an accurate prediction model.
- We have studied ML algorithms to see which one works well in predicting ASD.
- Finally, we have built a prediction model which will help in the early prediction of ASD in toddlers.

The remaining of this paper structured as follow: Section 2 discussed related works. The proposed methodology represented in Section 3. Section 4 illustrated the obtained results from the experiments. Conclusion and future work represented in Section 5.

2. Related Works:

Due to the increasing rates of ASD observed globally, many researchers have garnered attention to it. The prominent fields of research have been machine learning and data mining to predict ASD.

In early research work, Shuvo et al. [7] have built a model to predict ASD in adults using the Random Forest (RF) algorithm. The prediction was based on behavioral attributes in which they followed these steps to build the model. The first step was data collection, in which they used a dataset available online from the UCI repository. Secondly, data preprocessing by applying filters to clean the data, finally used a prediction algorithm in which they used RF and got 0.96% accuracy.

Raj and Masood [8] focused on using machine learning techniques and deep learning techniques to detect ASD at early stages. The researchers attempted 6 algorithms on 3 datasets for different ages of ASD. They started the work by doing pre-processing for the datasets to find any problem in the data that affected the detection process, then they applied Naïve Bayes, support vector machine (SVM), Convolutional Neural Network (CNN), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) on the datasets. The results they got showed that CNN was the best one for the detection process with accuracy: 99.53% for autistic adults, 98.30% for autistic children, and 96.88% for autistic adolescents.

Alwidian et al. [9] investigated association classification (AC) techniques to predict ASD in adults. They used a dataset for adults from the UCI repository and implemented the algorithms using WEKA software. The best AC algorithm was the Weighted Classification Based on Association Rules (WCBA) with 97% accuracy. Akter et al. [10] the authors used datasets for ASD from Kaggle and UCI repository, which are an open-source website for datasets. They chose datasets for toddlers, children, adolescents, and adults to help in the early detection of ASD. They applied ANN, recurrent neural network, decision tree, KNN, gradient boost, extrem learning machine, LR, Naïve Bayes, RF, SVM, and xgboost on the chosen datasets and made a pre-processing on data, then calculated the evaluation metrics for each of algorithms and compared between the applied algorithms. The LR gave the best result for all datasets, with accuracy of 100% for toddlers, 99.3% for children, 95% for adolescents and 99.8% for adults.

Nurisa et al. [11] introduced a model to predict attention deficit hyperactivity disorder (ADHD) using data mining. They used a dataset that contained data for ADHD children, then applied the mining approach by

doing data pre-processing. The modeling in which they built the model using regression and ANN. Evaluation in which they evaluated the performance, finally represented the knowledge and defined the signs that helped in predicting ADHD. The f1-score they got was 82.85%, which is better compared to previous research.

Jaafer et al. [12] proposed a model to detect autism spectrum disorder in adults. In this work, the researchers used an ASD dataset for adults with 21 attributes from the UCI Machine Learning repository and applied the following algorithms: logistic regression, sequential minimum optimization (SMO), Naïve Bayes, and instance-based technique based on k-neighbors (IBK) with Weka tools to classify the data for the adults. The results obtained showed that SMO has the highest accuracy, 99.71%.

Bala et al. [13] employed various machine learning models, including Naïve Bayes, K-star, Decision Tree (C4.5), Classification and Regression Trees (CART), KNN, SVM, Bagging Classifier (BG), and RF to detect ASD at early stage using toddlers, children, adolescent, and adult datasets. Their results indicated that SVM is the best one with accuracies of 96.67% for toddlers, 95.48%for children, 95.48% for adolescents, and 96.06% for adults. Another study, Khan et al. [14], has a review paper on research that discusses how to apply Machine Learning techniques to detect or predict ASD. In this paper, the researchers highlighted some lessons that need to be considered. Some of these lessons are: SVM, RF, and LR are the outperforming algorithms in ASD research, and precision and recall are the most used performance metrics. Rasul et al. [4] used Jupiter Notebook, which is Google's cloud-based service, to build a model that can diagnose ASD early. They used datasets for children and adults, and combined the children's and adults' datasets. The study used Naïve Bayes, KNN, LR, SVM, Decision Tree, RF, Boosting, and Artificial Neural Network (ANN), applying 8 performance metrics with100% accuracy for LR and SVM on children's and adult datasets, 94.24% accuracy for ANN on combined datasets. In a recent study, Chauhan et al. [15] combined ML with a database provider using the predictive method to predict ASD. They used SVM, KNN, RF, and decision tree. The best result got from RF with 74% accuracy.

R. S et al. [25], the authors developed a system that combined two modules to detect ASD in toddlers as early as possible. The first module called "Analysis sub-module" depending on behavioral data and used RF to analyze the data, the second called "Image sub-module" and used CNN to analyze the image data, the result from each module merged to define the final diagnosis and defined the severity levels of ASD on toddlers which provided robustness and reliability in the prediction process. The RF achieved an accuracy 85%, CNN achieved 87%, and the combination of two modules achieved 90%. Generalizability one of the limitations of this system as the data is collected from one country.

In the following Table 1, we represent the most popular datasets that are used in ASD research and the best algorithms that gave good results for building a prediction model. In our study, we focused on predicting ASD as early as possible in toddlers from 12 to 36 months by using datasets for ASD toddlers that we worked on and collected data from many hospitals, children's care centers, and clinics. Employing different machine learning techniques, we chose the best one to build a model that can predict ASD among toddlers.

Table 1 : Comparison between related works

Authors	Dataset	Target population	Best Algorithm	Accuracy

Shuvo et al. [7] (2019)	UCI repository	Adult	Random Forest	0.96%
Raj et al. [8] (2020)	UCI repository	Children, adult, adolescents	CNN	99.53% adult, 98.30 children, 96.88% adolescents
Alwidian et al. [9] (2020)	UCI repository	Adult	WCBA	97%
Akter et al. [10] (2021)	Kaggle, UCI repository	Toddler, children, adolescent, adult	LR	100% toddlers, 99.3% children, 95%, adolescent, 99.8% adult
Nurisa et al. [11] (2022)	ADHD dataset	ADHD children	ANN	82.85%
Jaafer et al. [12] (2022)	UCI repository	Adult	SMO	99.71%
Bala et al. [13] (2022)	Dr.Fadi Thabtah dataset	Toddlers, children, adolescent, and adult	SVM	96.67% toddlers, 95.48% children, 95.48% adolescent, 96.06% adult
Rasul et al. [4] (2024)	UCI repository	Children, adults	LR, SVM	100%
Chauhan et al. [15] (2024)	Not specified	Not specified	RF	74%

3. **The Proposed Methodology:**

Our study goal is to predict ASD in children at early stage as possible to help them get the required care that helps them to live a decent life. By understanding the ASD, we can enhance screening and diagnosis processes by applying machine learning algorithms to a dataset to identify patterns associated with ASD traits. Our methodology followed a structured pipeline, starting with data collection, preprocessing, then we trained and evaluated different models to assess their accuracy in the prediction process. We detail the key steps taken throughout this process:

3.1. Data Collection:

For this study, we worked on building a new dataset that includes information relevant to ASD. The data was gathered from Ain Shams Center for Special Needs Care, the Egyptian Autistic Society, Resala Charity Organization, and interviews with ASD parents. It contains 200 instances with 17 attributes. The attributes cover demographic data such as age and gender, answers to ASD test questions, and the final result diagnosis of whether a child has ASD or not.

The dataset is available publicly on the Science Data Bank under the title “Autism Spectrum Disorder Symptomatic dataset: For Arab Children” [16], to allow researchers to explore the ASD in the Arab community and develop prediction and diagnosis tools.

3.2. **Preparing the Data:**

During the data collection process, drops could occur, leading to inconsistencies in the data. Therefore, before applying machine learning algorithms, we should ensure that the data is ready for the analysis process to avoid any mistakes or problems that lead to wrong analysis. We start by handling the missing data using WEKA's ReplaceMissingValues filter to ensure the data completeness, then use a normalization filter to ensure that all numeric values in the dataset fall in the same range [0-1], and applying the remove filter and remove "first diagnosis attribute", and finally applying Discretize filter because some algorithms we used perform better with discrete and give better performance.

3.3. **Splitting Dataset:**

To effectively train the machine learning model and ensure its accuracy in making predictions, the dataset is split into training and testing sets using two approaches due to the dataset size (200 instances) and the algorithms that are used. These approaches are:

1. K-fold cross validation: in which the dataset is divided into equally sized K folds randomly. The K set is used for testing, and K-1 sets are used for training the model. The most popular values for K are K=5, and K=10 [17]. We used K=10 as it is more suitable to dataset size and gives better prediction and performance.
2. Train-test split: in which the dataset is divided into two sets: one for training the model and one for testing. The most popular splitting percentages are 80:20 and 70:30 due to the size of the dataset [18]. We choose a percentage of 70:30 to apply to our work.

3.4. Model Development:

We choose 5 machine learning algorithms to test them to determine which one gives the best accuracy and is effective in predicting ASD. The algorithms are as follows:

1. Naïve Bayes: it is an algorithm that is based on the Bayesian statistics theorem, which calculates probabilities by considering prior knowledge of conditions related to an event [19]. It is represented mathematically.

$$P(Y/X) = \frac{P(X \text{ and } Y)}{\mathrm{P(X)}} \quad \text{Eq.(1)}$$

2. Decision trees: used for classification and regression tasks. It is structured in a tree format which is root, branches and leaf. The leaf nodes represent the possible output within the dataset [20].

3. Support vector machine (SVM): it is commonly used in health research due to its high efficiency and robustness. SVM works by finding the optimal hyperplane for the prediction model and making it accurate [21].

4. Logistic Regression (LR): it is used for classification purposes, in which the output value belongs to the probability of a specific class.[22]

5. Artificial Neural Network (ANN): It is an algorithm that uses the input and output data to find the factors that lead to output data. It can train on new data which makes the prediction process more accurate. It is like the human brain functions that it is able to receive many inputs, process them, and give the output value [23].

4. Experiment Result:

To validate the quality and applicability of the "Autism Spectrum Disorder Symptomatic dataset: For Arab Children", we conducted a series of ML experiments using several ML algorithms. The aim of these experiments is to measure the ability of our dataset to support research that aims to predict ASD early.

We carried out our experiments using WEKA 3.9 [24], a widely used open-source machine learning platform. The dataset contains 200 samples with 17 attributes, including demographic data, response to ASD screening questions, and diagnosis label ASD, or not.

We applied 10- fold cross-validation and train testing split to ensure robustness and get more accurate and reliable results. The following tables represent the results we got from the 10-fold cross-validation and train-test split. Table 2 represents the accuracy and other performance metrics for the selected ML algorithms, the results based on a 10-fold cross-validation technique, and Table 3 presents the performance metrics for ML algorithms based on training-testing techniques. In addition, Figure 1 shows the results in a graph and indicates that our dataset can be a practical tool for building prediction models that use ML algorithms, especially in Arab populations, based on any splitting techniques

Table 2: 10-fold cross-validation accuracy and performance metrics

Algorithms	Accuracy	Precision	Recall	F-Measure
Naïve Bayes	94.92%	95%	94%	95%
Decision Trees	87.81%	87%	87%	87%
SVM	94.41%	94%	94%	94%
LR	92.38%	92%	92%	92%
ANN	96.44%	96%	96%	96%

Table 3 : Training-testing splitting accuracy and performance metrics

Algorithms	Accuracy	Precision	Recall	F-Measure
Naïve Bayes	93.22%	93%	93%	93%
Decision Trees	88.13%	88%	88%	88%
SVM	91.52%	91%	91%	91%
LR	89.83%	90%	89%	89%
ANN	91.52%	91%	91%	91%

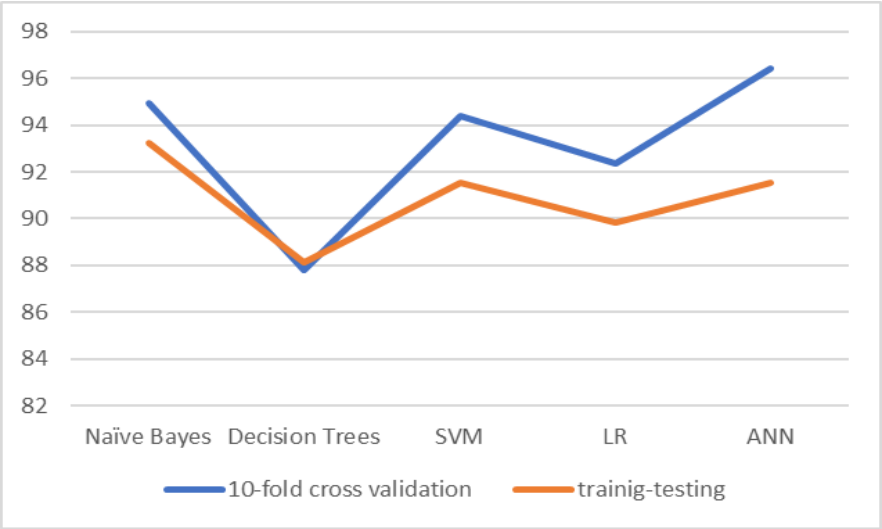


Figure 1 Splitting dataset approaches results

5. Conclusion and Future Work:

In this research, we used our dataset "Autism Spectrum Disorder Symptomatic dataset: For Arab Children" to develop a prediction model that can predict ASD in toddlers at an early stage. We applied 5 ML algorithms Naïve bayes, Decision trees, SVM, LR, and ANN, using 2 approaches to splitting data 10-fold cross-validation and training-testing methods, after made a pre-processing for the data to ensure that is suitable to feed it to the model. These methods enabled us to find the best performance metrics and assess model stability. The findings indicate that the ANN model outperformed other algorithms under the 10-fold cross-validation achieving the highest accuracy 96.44%, and Naïve Bayes delivered the best in a training-testing approach with accuracy 93.22%. These findings demonstrate the importance of developing ML model to predict ASD on toddlers, especially within specific cultural populations such as Arab population.

For future work, the researchers can complete the dataset to make it wider and include many Arab regions and populations, so the dataset can become a stepping stone for future work related to ASD by encouraging more research, fostering collaboration, and contributing to a more accessible and early diagnosis of ASD.

6. References:

[1]"Autism.": https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders , (Accessed: Feb. 20, 2025).
[2] R. L. Simpson, S.R. de Boer-Ott, D.E.Griswold, B.S.Myless, S.E.Byrd, J.B.Ganz, K.T.Cook, K.L.Otten, J.Ben-Arieh, S.Kline, and L.G.Adams "Autism Spectrum Disorders: Interventions and Treatments for Children and Youth". 2005.
[3]"Autism Spectrum Disorder - National Institute of Mental Health (NIMH)." https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd, (Accessed: Feb. 21, 2025).

[4] R. A. Rasul, P. Saha, D. Bala, S. M. R. U. Karim, M. I. Abdullah, and B. Saha, "An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder," Healthcare Analytics, vol. 5, Jun. 2024, doi: 10.1016/j.health.2023.100293.
[5] J. L. Matson and J. L. Matson Editor, "Autism and Child Psychopathology Series Series Editor: Handbook of Assessment and Diagnosis of Autism Spectrum Disorder.
[6] F. Thabtah, R. Spencer, N. Abdelhamid, F. Kamalov, C. Wentzel, Y. Ye, and T. Dayara "Autism screening: an unsupervised machine learning approach," Health Information Science and Systems, vol. 10, no. 1, Dec. 2022, doi: 10.1007/s13755-022-00191-x.
[7] S. B. Shuvo, J. Ghosh, and A. S. Oyshi, "A Data Mining Based Approach to Predict Autism Spectrum Disorder Considering Behavioral Attributes," 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, Jul. 2019, doi: 10.1109/ICCCNT45670.2019.8944905.
[8] S. Raj and S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," in Procedia Computer Science, Elsevier B.V., 2020, pp. 994–1004. doi: 10.1016/j.procs.2020.03.399.
[9] J. Alwidian, A. Elhassan, and R. Ghnemat, "Predicting Autism Spectrum Disorder using Machine Learning Technique," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 5, pp. 4139–4143, Jan. 2020, doi: 10.35940/ijrte.E6016.018520.
[10] 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 International Conference on Robotics, Electrical and Signal Processing Techniques, 2021, pp. 742–747. doi: 10.1109/ICREST51555.2021.9331013.
[11] M. Nurisa, G. Dastghaibyfard, and H. Hadianfard, "Predicting ADHD from early childhood data using data mining." Apr. 25, 2022. doi: 10.21203/rs.3.rs-1395357/v1.
[12] S. Jaffer, I. Abdulazez, N. Al-Qazzaz, and T. Yousif, "Data Mining for Autism Spectrum Disorder detection among Adults," Al-Nahrain Journal for Engineering Sciences, vol. 25, no. 4, pp. 142–151, Dec. 2022, doi: 10.29194/njes.25040142.
[13] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, "Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder," Algorithms, vol. 15, no. 5, May 2022, doi: 10.3390/a15050166.
[14] K. Khan and R. Katarya, "Machine Learning Techniques for Autism Spectrum Disorder: Current trends and future directions," in 2023 International Conference on Innovative Trends in Information Technology, ICITIIT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICITIIT57246.2023.10068658.
[15] R. Chauhan, K. Mehta, Y. Eiad, and M. F. Zuhairi, "Prediction of Autism Spectrum Disorder Using AI and Machine Learning," in Proceedings of the 2024 18th International Conference on Ubiquitous Information Management and Communication, IMCOM 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IMCOM60618.2024.10418312.
[16] Dina Ayman. Autism Spectrum Disorder Symptomatic dataset: For Arab Children [DS/OL]. V2. Science Data Bank, 2025[2025-04-20]. https://cstr.cn/31253.11.sciencedb.20788. CSTR:31253.11. sciencedb.20788.
[17] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran, and A. Abraham, "A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools," Engineering Applications of Artificial Intelligence, vol. 111, p. 104773, May 2022, doi: 10.1016/J.ENGAPPAI.2022.104773.

[18] "Splitting Data for Machine Learning Models | GeeksforGeeks." https://www.geeksforgeeks.org/splitting-data-for-machine-learning-models/ , (Accessed: Mar. 13, 2025).
[19]"What Are Naïve Bayes Classifiers? | IBM." https://www.ibm.com/think/topics/naive-bayes , (Accessed: Mar. 18, 2025)
[20]A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2017.
[21]"What is a Decision Tree? | IBM." https://www.ibm.com/think/topics/decision-trees, (Accessed: Mar. 19, 2025).
[22] G. Khyathi, K.P. Indumathi, J. A, L.F.Jency M, S.Siluvai, and G. Krishnaprakash, "Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health". Cureus, 17(1): e77169. (2025) doi:10.7759/cureus.77169
[23]"How Do Neural Networks Work? Your 2025 Guide | Coursera." Accessed: Mar. 19, 2025. [Online]. Available: https://www.coursera.org/articles/how-do-neural-networks-work
[24]"The Data Platform for Cloud & AI | WEKA - WEKA."]. https://www.weka.io/, (Accessed: Apr. 20, 2025).
[25] R. S.C., S. Cirillo, Y. D., and L. Solimando, "A hybrid approach combining images and questionnaires for early detection and severity assessment of Autism Spectrum Disorder," *Image and Vision Computing*, vol. 160, Jul. 2025, doi: 10.1016/j.imavis.2025.105547.