Article

Innovative Approach to Detecting Autism Spectrum Disorder Using Explainable Features and Smart Web Application

Mohammad Abu Tareq Rony [1,†], Fatama Tuz Johora [2,3], Nisrean Thalji [4], Ali Raza [5,†], Norma Latif Fitriyani [6,†], Muhammad Syafrudin [6,*] and Seung Won Lee [7,8,9,10,*]

1   Department of Statistics, Noakhali Science & Technology University, Noakhali 3814, Bangladesh;
     rony1513@student.nstu.edu.bd
2   Department of Computer Science and Engineering, University of Chittagong, Chittagong 4331, Bangladesh;
     fatama@aintlab.com
3   Applied INTelligence Lab (AINTLab), Seoul 05006, Republic of Korea
4   Faculty of Computer Studies, Arab Open University, Amman 11953, Jordan; n_thalji@aou.edu.jo
5   Department of Software Engineering, University of Lahore, Lahore 54000, Pakistan; ali.raza@se.uol.edu.pk
6   Department of Artificial Intelligence and Data Science, Sejong University, Seoul 05006, Republic of Korea;
     norma@sejong.ac.kr
7   Department of Precision Medicine, Sungkyunkwan University School of Medicine,
     Suwon 16419, Republic of Korea
8   Department of Metabiohealth, Sungkyunkwan University, Suwon 16419, Republic of Korea
9   Personalized Cancer Immunotherapy Research Center, Sungkyunkwan University School of Medicine,
     Suwon 16419, Republic of Korea
10   Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, Republic of Korea
*   Correspondence: udin@sejong.ac.kr (M.S.); swleemd@g.skku.edu (S.W.L.)
†   These authors contributed equally to this work.

Abstract: Autism Spectrum Disorder (ASD) is a complex developmental condition marked by challenges in social interaction, communication, and behavior, often involving restricted interests and repetitive actions. The diversity in symptoms and skill profiles across individuals creates a diagnostic landscape that requires a multifaceted approach for accurate understanding and intervention. This study employed advanced machine-learning techniques to enhance the accuracy and reliability of ASD diagnosis. We used a standard dataset comprising 1054 patient samples and 20 variables. The research methodology involved rigorous preprocessing, including selecting key variables through data mining (DM) visualization techniques including Chi-Square tests, analysis of variance, and correlation analysis, along with outlier removal to ensure robust model performance. The proposed DM and logistic regression (LR) with Shapley Additive exPlanations (DMLRS) model achieved the highest accuracy at 99%, outperforming state-of-the-art methods. eXplainable artificial intelligence was incorporated using Shapley Additive exPlanations to enhance interpretability. The model was compared with other approaches, including XGBoost, Deep Models with Residual Connections and Ensemble (DMRCE), and fast lightweight automated machine learning systems. Each method was fine-tuned, and performance was verified using k-fold cross-validation. In addition, a real-time web application was developed that integrates the DMLRS model with the Django framework for ASD diagnosis. This app represents a significant advancement in medical informatics, offering a practical, user-friendly, and innovative solution for early detection and diagnosis.

Keywords: autism; data mining; machine learning; ANOVA; logistics regression; web app

MSC: 68T01; 62J10; 68U35; 68T07

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that inhibits the typical maturation of essential communication and social functions [1]. This collection

of mental health conditions disrupts normal brain development, resulting in challenges in social and communication skills [2]. The term "autism spectrum" reflects the wide range of autism manifestations among individuals, as each person with autism experiences it uniquely. Consequently, ASD is described as a "spectrum disorder", with varying support needs among individuals to achieve their desired quality of life [3]. Genetic and neurological factors are also associated with ASD. Although ASD has genetic foundations, it is primarily diagnosed based on behavioral markers including social interaction, creativity, repetitive behaviors, and communication. Co-occurring conditions, such as epilepsy, depression, anxiety, attention-deficit hyperactivity disorder, sleep disturbances, and self-injurious behaviors, are commonly observed in individuals with autism.

Intellectual function in people with autism varies significantly, ranging from severe disability to exceptional intellectual abilities [4]. Research suggests that both environmental and genetic factors may increase a child's susceptibility to autism. For instance, if an older sibling is on the spectrum, a child's likelihood of having autism increases by approximately 19% [5]. Emerging evidence points to maternal infections, certain medications, and parental age (particularly advanced paternal age) as potential risk factors [5]. When one child in a family is diagnosed with ASD, the risk for subsequent children increases by 20%, and the chance of ASD rises by approximately 32% if the first child has ASD [6].

Genetic causes are identifiable in approximately 10–20% of ASD cases. Autism is considered a "spectrum" disorder because the type and severity of symptoms vary widely across affected individuals [7]. Although autism can be diagnosed at any age, it is classified as a "developmental disorder" since symptoms often appear within the first two years of life. Individuals with autism have complicated healthcare requirements that require integrated treatments encompassing health promotion, medical care, and rehabilitation. Collaboration across healthcare, education, employment, and social services is essential. According to one study, approximately 33% of children with other developmental conditions show some ASD symptoms but do not meet the full diagnostic criteria [8].

There are several clinical and nonclinical methods for diagnosing ASD. Clinical diagnostic techniques include the Autism Diagnostic Observation Schedule-Revised and the Autism Diagnostic Interview [9]. Most current ASD diagnostic methods require significant time to complete. Recently, researchers have begun to integrate machine learning (ML) technologies to streamline ASD diagnosis. The primary objectives of ML studies on ASD are to reduce the dimensionality of input datasets for identifying the most relevant ASD features, improve diagnostic accuracy, and decrease the time needed for diagnosis, thereby facilitating faster access to healthcare services. DM is a field that integrates mathematics, artificial intelligence (AI), search algorithms, and other scientific disciplines to develop reliable predictive models from autism datasets [10].

Early identification of autism can be beneficial for children by providing targeted assistance to meet their unique needs [3]. This project used ML techniques to analyze ASD across diverse populations globally. Additionally, the Quantitative Checklist for Autism (Q-Chat) and other variables were included in the ASD test application. We developed a simple model to estimate the probability of ASD traits, allowing parents to take early action. Exploratory data analysis helped identify essential factors related to autism. Moreover, identifying trends and patterns through statistical DM significantly impacts ASD detection. The growing prevalence of ASD worldwide, along with its social and economic implications, underscores the significance of developing efficient and practical screening procedures.

ASD affects approximately 1% of the global population (about 62.2 million as of 2015) [11], with a higher diagnosis rate in males than females [12]. Although medications can aid in managing symptoms, they offer limited long-term benefits. Thus, early ASD detection is crucial. Early diagnosis enables timely intervention, fostering improved social, cognitive, and communication skills. Addressing developmental needs early in life contributes to better long-term outcomes by mitigating symptoms and improving overall development in individuals with ASD.

The primary contributions of this research work are as follows:

- Despite extensive research on ASD, this study is the first to identify relevant features through ANOVA and Chi-Square analyses and to examine possible correlations before fitting the proposed DMLRS model, enhancing prediction accuracy.
- Secondary data were collected using a mobile application for research, incorporating ten research questions (A1–A10) and information on variables including age, jaundice history, ethnicity, sex, prior app usage, family relationships, ASD presence in family members, and the dependent variable: autism classification.
- We developed an innovative autism prediction model integrating XAI with ML and DM algorithms, achieving higher predictive performance. A comparative analysis with state-of-the-art models is also provided.
- Finally, this study implemented the proposed model in a web application featuring a user-friendly interface to support individuals and healthcare providers in assessing autism.

This study is organized as follows: Section 2 presents an analysis of the related literature. In Section 3, we describe the proposed model for ASD detection. Section 4 compares the results obtained from implementing various techniques. Section 4.6 details the web application system, and finally, Section 5 summarizes the findings of this research study.

2. Literature Review

The literature analysis section establishes the contextual framework for ML applications in pediatric ASD diagnosis. Direct interactions with medical professionals are essential in managing ASD in children, involving a comprehensive evaluation of the child's developmental history, responsiveness, behavioral patterns, attention capabilities, and Intelligence Quotient. Typically, children with ASD begin to exhibit specific symptoms around the age of three, such as sensory sensitivity, speech and communication challenges, coordination difficulties, and notable changes in emotional and social well-being. This section offers a detailed review of ML and deep learning (DL) methodologies employed for diagnosing ASD from images or numerical datasets, complemented by a comparative analysis provided in with a comparative analysis summarized in Table 1.

Table 1. Summary of literature on ASD detection and analysis.

Ref.	Problem Statement	Research Objective	Main Contribution	Experimental Result
[13]	Analysis and identification of ASD.	Early diagnosis of ASD.	Detecting ASD and analyzing ASD issues.	The Neural Network-based model may identify ASD instead of a typical DM classifier.
[14]	Behavioral research on ASDs.	Incorporate an intelligent DM algorithm into an existing diagnostic tool.	Surpasses the majority of previous studies.	SVM (Support Vector Machine) is used to generate ASD classification models.
[10]	Autism detection based on rule induction.	To increase the efficiency of ASD identification.	Rule-based representations of automatic classification systems.	Rule-based superior algorithms that offer higher sensitivity rates.
[15]	Evaluation of ASD DM classification.	Create a different model with a greater capacity for early predicted ASD.	Using Autism Questions (AQs) to create models.	LR achieved the maximum accuracy using the Chi-Square approach.
[16]	ASD diagnosis employing optimal techniques.	Expedite autism diagnosis.	Predict a person's ASD symptoms and identify the most effective model.	Provide good experimental results.
[17]	Examine Q-Chat for early autism screening.	Investigation of the accuracy and reliability of the quantitative autism screening tool for toddlers, called Q-Chat.	Demonstrates exceptional accuracy, showcasing the tool's robust performance and cross-cultural reliability.	SVM proved to be an effective classification method.

Table 1. *Cont.*

Ref.	Problem Statement	Research Objective	Main Contribution	Experimental Result
[18]	Screening for ASD using machine learning models.	Prediction of ASD to facilitate diagnosis and subsequent treatment.	Prediction of ASDs.	On the adult autism dataset, a neural network has the highest accuracy.
[19]	Models that are based on DM for the early detection of ASD.	Early detection of ASD.	Introduced a DM-based model that may be applied to the early detection of ASD.	SVM gives a better result.
[20]	Applying DM techniques to predict ASD.	Develop and use a model for the early prediction of autism.	Proposed a DM approach to conducting early prediction of ASD.	Greatest accuracy among other study disciplines.
[21]	Predicting autistic features by replacing conventional scoring systems.	Designing an accurate screening system for autism.	Increasing the screening process's accuracy.	CNN is the best algorithm for detecting ASD features compared to DM methods.
[22]	ASD detection.	Identify the most important characteristics and automate the diagnostic process aim of improving diagnosis.	Analyzing the features of ASD datasets and finding correlations.	Neural network classifier beats all other benchmark DM algorithms.
[23]	Identification of ASD in children.	Assess if a child is prone to ASD in its earliest stages, streamlining the process of diagnosis.	Proposed a predictive model with the highest accuracy to identify ASD in children.	LR provides the greatest accuracy.

ASD is a complex neurodevelopmental disorder characterized by persistent challenges in social communication and interaction, as well as restrictive and repetitive behaviors, interests, and activities. Early detection and intervention are crucial to optimize outcomes for individuals with ASD, enabling tailored support to address unique needs. Recent advancements in technology and data analysis have spurred catalyzed research focused on improving the detection and analysis of ASD, with studies employing various methodologies and innovative approaches.

A notable contribution to the field is presented by Raj [13], who proposed a neural-network-based model for early ASD diagnosis. Leveraging ML techniques, Raj's model demonstrated strong performance in identifying ASD cases, offering potential advantages over traditional diagnostic methods. This novel approach shows significant promise for improving the accuracy and efficiency of ASD diagnosis, particularly in early childhood when timely intervention is most beneficial.

In a similar vein, Hriti [24] conducted research on ASD diagnosis by integrating visual and behavioral data from ASD patients and neurotypical individuals. By combining multiple data modalities—including visual cues and behavioral patterns—Hriti's study highlighted the superiority of multi-modal data integration in enhancing ASD detection accuracy. This approach provides a more comprehensive understanding of ASD and highlights the importance of incorporating diverse data sources in the diagnostic process.

Thabtah et al. [14] made significant contributions to ASD classification by incorporating Support Vector Machines (SVMs) and rule-based algorithms, demonstrating the effectiveness of DM algorithms in surpassing previous performance benchmarks. Their research highlighted how advanced computational techniques could enhance ASD diagnosis, particularly in clinical settings where early detection is essential for intervention planning.

In related advancements, Abdullah [15] proposed using Autism Questions to improve early ASD prediction models. Logistic regression (LR) emerged as the most accurate model in Abdullah's study, achieving maximum accuracy through the Chi-Square approach. This highlights the importance of feature selection and model optimization in refining ASD detection algorithms.

Similarly, Alteneiji [16], Tartarisco [17], and Baranwal [18] developed predictive models and screening tools to facilitate early ASD detection and treatment. Their studies aimed to bridge the gap between diagnosis and intervention, ultimately enhancing outcomes for individuals with ASD.

Furthermore, Akter [19], Musa [20], and Shahamiri [21] proposed DM-based models for early ASD detection, with Convolutional Neural Networks (CNNs) exhibiting superior performance in detecting ASD features compared to traditional DM techniques. Their research underscores the potential of DL in improving diagnostic accuracy.

Hossain [22] analyzed ASD dataset features across age groups, identifying correlations between specific traits and ASD diagnoses. This work deepens understanding of ASD's fundamental mechanisms and contributes to the development of precise diagnostic tools.

Vakadkar [23] introduced a predictive model tailored to identify ASD in children, with LR achieving the highest accuracy among classification algorithms. This research is significant for early intervention programs, as the accurate, timely identification of ASD in children is crucial for accessing targeted support services.

Joudar et al. [25] addressed the need for advanced AI-based diagnostic methods for ASD due to its complexity and widespread public concern. Their study systematically reviews AI applications in early ASD diagnosis and triage, analyzing 46 recent studies that enhance diagnostic accuracy and identify areas for future research. This work underscores AI's expanding role in ASD healthcare, proposing the use of AI and fuzzy Multi-Criteria Decision Making (MCDM) methods to improve patient triage and prioritization. The proposed methodology is structured into five distinct phases, bridging theoretical concepts with practical applications.

Similarly, Albahri et al. [26] proposed an explainable AI framework for ASD triage using fuzzy MCDM, aimed at efficiently categorizing ASD severity based on diverse data inputs. This framework encompasses five phases and introduces four novel algorithms to classify patients into three severity levels. Experimental results, obtained from balanced ASD datasets and two AI models, demonstrated the framework's efficacy in balancing and interpreting data, suggesting its potential for clinical utility.

Building on these advancements, Joudar et al. [27] developed a new triage method for ASD using Fuzzy-MCDM (fMCDM), focusing on varying symptom presentations and severity levels. The methodology involved preprocessing an ASD dataset of 988 patients and implementing two fMCDM methods to prioritize influential criteria, resulting in processes for triaging patients with autism (PTAP). This method accurately triaged 538 patients into the minor, moderate, and urgent categories, demonstrating its effectiveness through sensitivity and specificity analyses. This approach supports early ASD diagnosis and treatment, significantly outperforming previous methods in comparative assessments, and provides a basis for future enhancements.

Finally, Joudar et al. [28] introduced a taxonomy for ASD triage and prioritization, using AI to construct a framework that simplifies the diagnostic process and identifies five major open issues in ASD triage. This research involved a systematic review of AI methodologies, examining 363 articles from sources such as ScienceDirect and PubMed, with a focus on diagnostic approaches, risky genes, and e-triage. The findings suggest a conceptual framework employing MCDM techniques to prioritize ASD patients by severity, aimed at enhancing diagnostic accuracy and patient care.

In summary, recent advancements in ASD detection and analysis have led to the development of innovative methodologies and tools designed to improve diagnostic accuracy and facilitate early intervention. Utilizing ML techniques, multi-modal data integration, and algorithmic innovations, researchers are addressing the complexities of ASD diagnosis and treatment. Collectively, these studies significantly advance the field of ASD detection and analysis by providing critical insights and novel approaches that enhance outcomes for individuals with ASD.

Research Gap and Questions

In the field of early ASD detection, the literature review identified a significant research gap, particularly in achieving accuracy, scalability, adaptability, computational efficiency, and real-time applicability in autism-detection methods. Our study specifically addresses two primary research questions arising from these gaps:

- Does employing a novel DMLRS technique—one that integrates prominent ML algorithms with DM techniques—improve the accuracy of ASD detection compared to existing methods?
- What are the most effective methodologies in ML, data mining, and web application development for accurately identifying autism?

To bridge these questions, we introduced an advanced technique aimed at improving the accuracy and efficiency of ASD detection. By integrating diverse ML and DM techniques, the proposed DMLRS offers a novel and robust solution for the evolving landscape of ASD detection.

3. Proposed Methodology

In Figure 1, the research methodology is presented in distinct subsections, covering key components such as an overview of the dataset, data preprocessing techniques, statistical analyses, logistic regression, and tree-based algorithms. Furthermore, Figure 1 provides a comprehensive illustration of the methodology, detailing the data preprocessing steps, which include missing value analysis and outlier removal via boxplot methods, as well as exploratory visualizations. It also showcases an array of XAI-based DM and ML algorithms assessed with different evaluation metrics. This project leverages RStudio and Python for their strengths in statistical computing, ML, and visualization within an open-source IDE. The final phase of the project involves developing a web application using the Django framework and performing a comparative analysis with related studies in ASD detection.
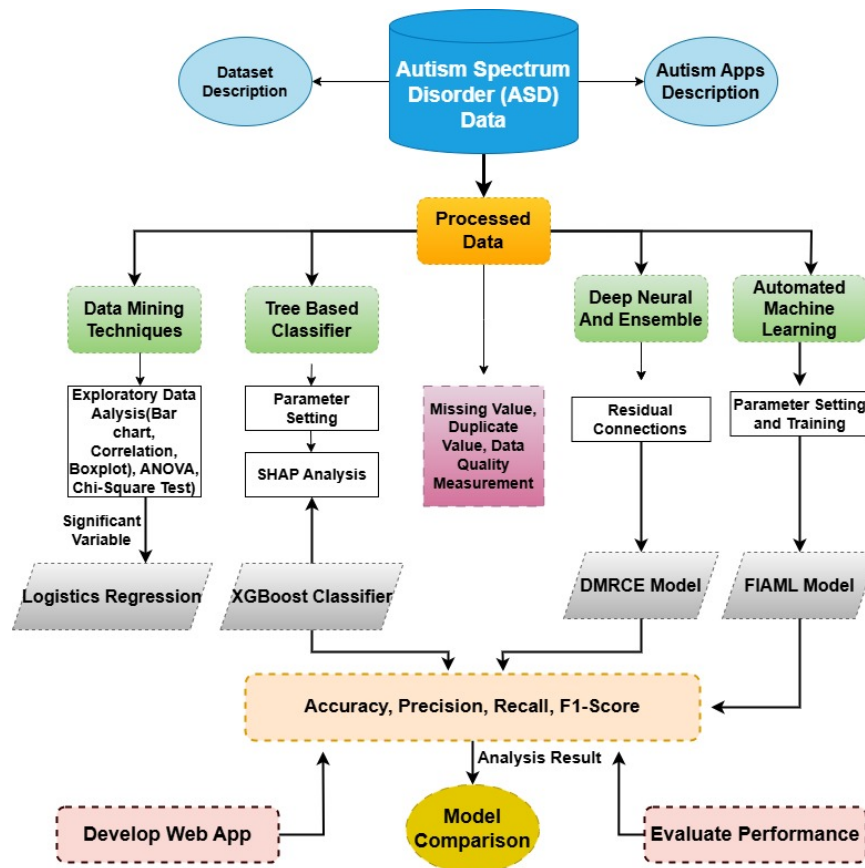

Figure 1. Proposed DMLRS methodology for ASD detection.

Algorithm 1 provides each step of our proposed method. The algorithm starts working from data preprocessing to model training, feature selection, and performance validation for ASD detection.

Algorithm 1 DMLRS Model

1: **Input:** Cleaned dataset
2: **Output:** Trained DMLRS Model
3: **procedure** DMLRS MODEL
4:    Load the cleaned trained dataset
5:    Perform feature selection using Chi-Square and ANOVA tests
6:       a. Select significant features (p-value < 0.05)
7:    Initialize LR model
8:    Train LR model using the selected features
9:       a. Split data into training and validation sets (e.g., 80:20)
10:      b. Use k-fold cross-validation (e.g., k = 10) for training
11:      c. Evaluate model performance using accuracy, precision, recall, and F1-score
12:   **if** model performance is satisfactory (e.g., accuracy $> 99\%$) **then**
13:         a. Save the trained LR model
14:   Initialize SHAP (Shapley Additive Explanations) for model interpretability
15:   Calculate SHAP values for the trained LR model
16:      a. Identify important features contributing to the model's predictions
17:   Return the trained LR model and SHAP values

3.1. Autism Spectrum Dataset and Preprocessing

The dataset of 1054 patient samples captures key demographic and clinical characteristics relevant to ASD, including age, gender, ethnicity, and core behavioral markers [29]. Although it does not represent all global variations, this dataset provides a broad and diverse sample that closely reflects commonly observed ASD features, thereby supporting reliable model generalization.

There were no missing data. However, age outliers were identified and subsequently removed using boxplot analysis. Table 2 summarizes the ASD dataset, while Table 3 lists five representative data samples. The experiment utilized publicly available autism data [30] previously employed in research aimed at enhancing autism prediction. The 20 selected variables focus on key behavioral and clinical indicators relevant to ASD diagnosis.

Table 2. Variables and their descriptions.

No	Variable Name	Variable Type	Variable Description
		Independent Variables	
1	Case No	Numeric	The participant's ID number.
2	A1	Binary (0, 1)	Is your child responsive when you call their name?
3	A2	Binary	How comfortable are you in establishing eye contact with your child?
4	A3	Binary	Does your child use pointing gestures to express their desires or needs?
5	A4	Binary	Does your child engage in pointing gestures to express shared interests with you?
6	A5	Binary	Does your child engage in pretend play, such as taking care of dolls or pretending to talk on a toy phone?

Table 2. *Cont.*

No	Variable Name	Variable Type	Variable Description
7	A6	Binary	Does your child track or follow your gaze direction?
8	A7	Binary	When you or someone else in the family is visibly upset, does your child display signs of wanting to offer comfort or consolation?
9	A8	Binary	Would you describe your child's first words as typical?
10	A9	Binary	Does your child use simple gestures?
11	A10	Binary	Does your child stare blankly or without reason?
12	Q-Chat Score	Numeric	The Q-CHAT score is a screening measure based on a 10-item (A1–A10) screening tool for autism in toddlers (18–24 months). Higher scores indicate a greater likelihood of autism, suggesting the need for further evaluation.
13	Age	Number	Age in months.
14	Sex	String	Gender.
15	Ethnicity	String	Ethnicities.
16	Jaundice	Boolean (Yes or No)	Jaundiced at birth.
17	Family member with ASD	Boolean	A family member has an ASD.
18	Relation	String	Relation to the child (e.g., Parent, Self, etc.).
19	Used app before	Boolean	Whether the participant has used this app before.
		Dependent Variables	
20	Class	Boolean	Participant classification as ASD or not ASD.

Table 3. Snapshot of 5 sample data.

Variable	Case_No 1	Case_No 2	Case_No 3	Case_No 4	Case_No 5
A1	0	1	1	1	1
A2	0	1	0	1	1
A3	0	0	0	1	0
A4	0	0	0	1	1
A5	0	0	0	1	1
A6	0	1	0	1	1
A7	1	1	1	1	1
A8	1	0	1	1	1
A9	0	0	0	1	1
A10	1	0	1	1	1
Q-Chat_score	3	4	4	10	9
Age (month)	18.605	13.829	14.679	61.035	14.256
Sex	f	m	m	m	f

Table 3. *Cont.*

Variable	Case_No 1	Case_No 2	Case_No 3	Case_No 4	Case_No 5
Ethnicity	Middle Eastern (ME)	White (WE) European	ME	Hispanic	WE
Jaundice	yes	yes	yes	no	no
Family mem ASD.	no	no	no	no	yes
Relation	family member (FM)	FM	FM	FM	FM
Used app before	no	no	no	no	no
Class	No	Yes	Yes	Yes	Yes

3.2. Autism Mobile App for Data Collection

The mobile application used for data collection is available on the Play Store [31]. Figure 2 illustrates the architecture of an AI-driven autism-detection system. This system comprises a mobile app, an intelligent web service that facilitates communication between the Autism AI app and ML models, a database to store user responses and test outcomes, and a screening algorithm designed for autism detection. The Autism AI application interacts with web services to utilize and deploy models. Its primary function is to offer a user-friendly interface for caregivers and family members, offering quick assessments of autistic traits. In addition, the app collects and validates critical user data, including behavioral patterns and demographic information.
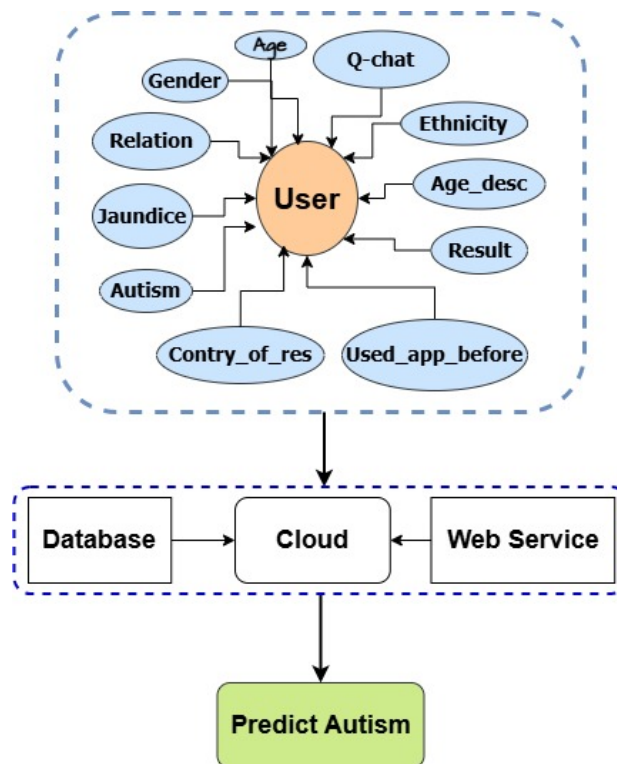

Figure 2. Architecture of the AI-driven autism-detection application.

3.3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) uses data visualization techniques to systematically explore and reveal essential dataset characteristics, allowing researchers to identify critical patterns and insights [32]. Serving as a foundational step in data analysis, EDA not only enhances understanding of the dataset's inherent traits but also informs the choice of suitable statistical methods [33].

Figure 3 presents a bar chart that highlights several key findings from the dataset. A notable portion of the sample population shows a high risk for autism, with most participants originating from North America. Additionally, a substantial number of participants have a history of jaundice. Ethnic analysis reveals a predominance of South Asian, Middle Eastern, and White European individuals, with an approximately equal gender distribution (male/female) across the sample.
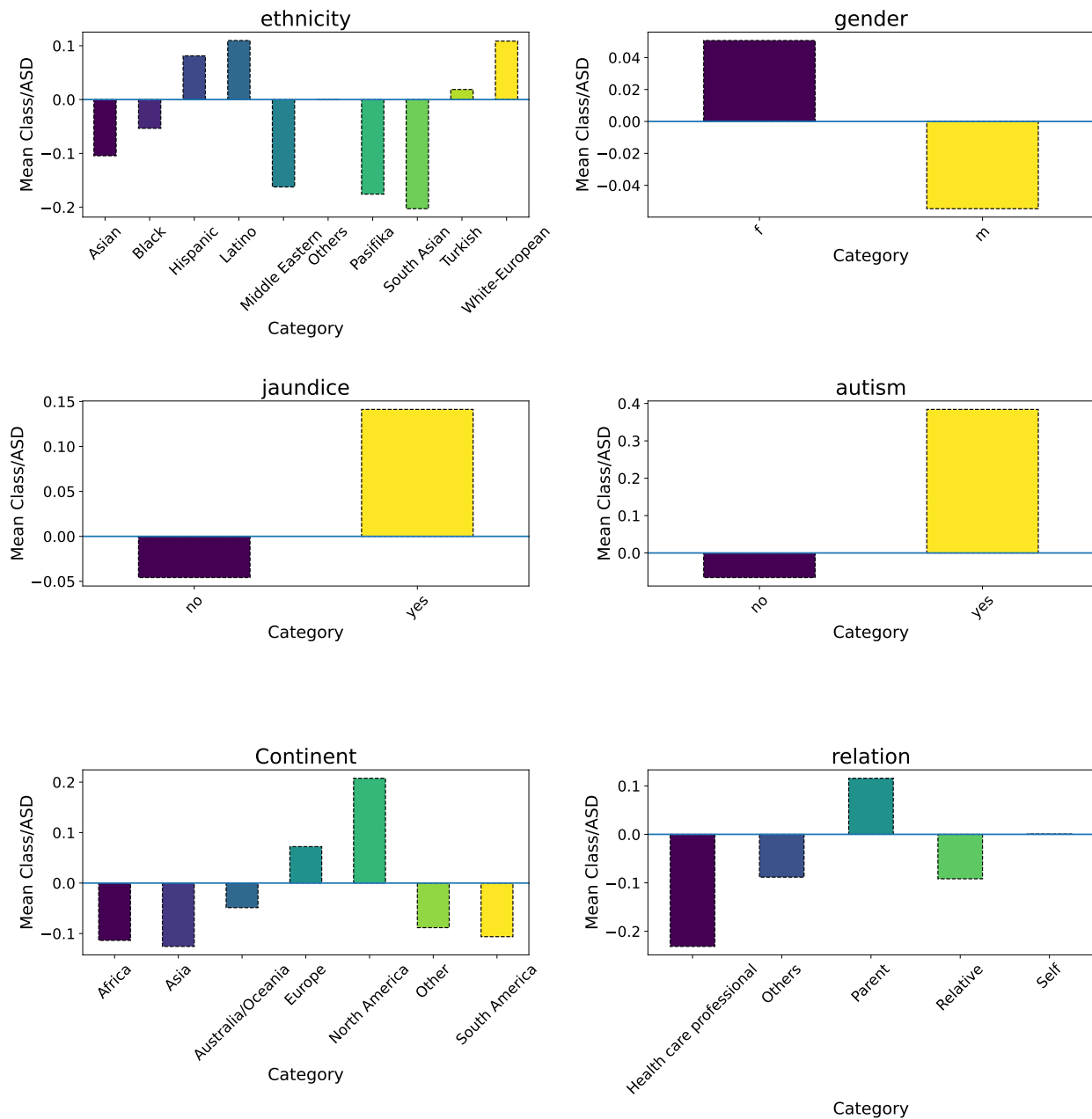

Figure 3. Bar chart of categorical variables.

Figure 4 presents a bar plot in which each rectangular bar displays statistical information proportional to the frequency of responses in the autism dataset. The plot visually represents the values of the A1–A10 variables, where the length of each bar corresponds to the number of responses. Notably, for each variable, a "Yes" response is associated with a higher likelihood of an autism diagnosis.

As shown in Table 4, the *p*-value is below the significance level for variables including sex, ethnicity, jaundice, and questions A1–A8. This result leads to the rejection of the null

hypothesis, indicating that these variables have a statistically significant relationship with the dependent variable class. Thus, these variables should be prioritized in subsequent analyses. Conversely, variables such as family members with ASD, relation, previous app use, and questions A9 and A10 did not show significance and are therefore excluded from further analysis.
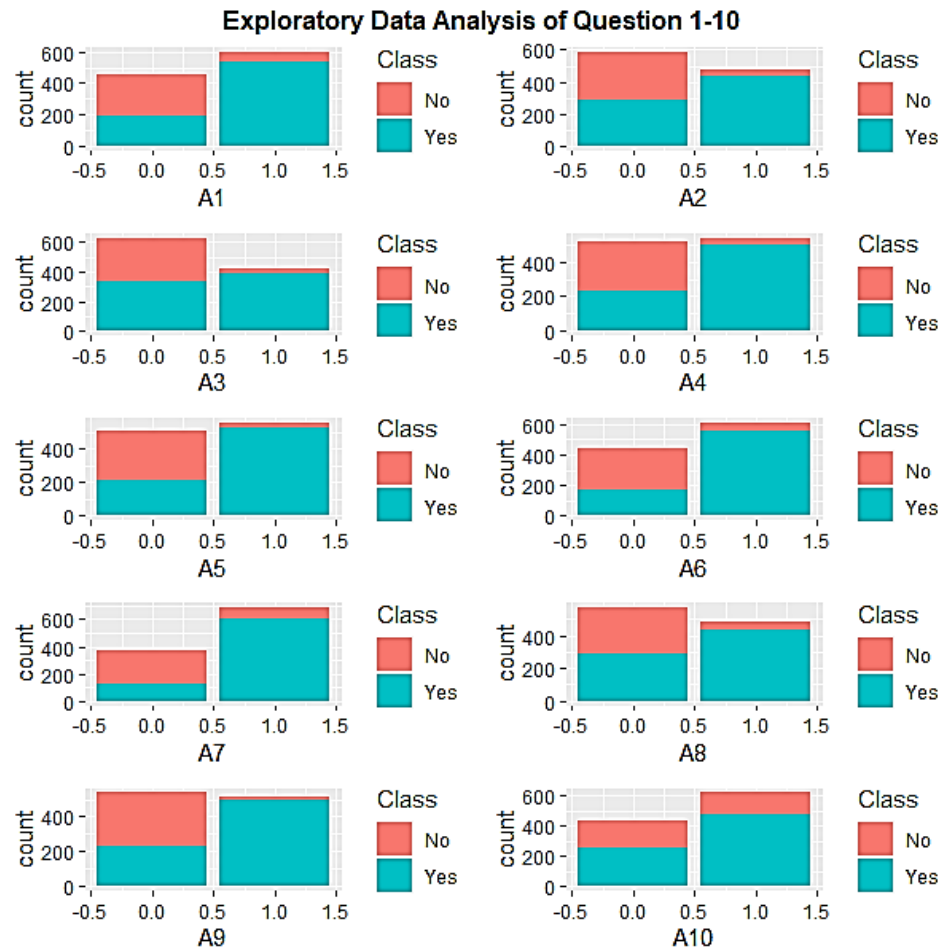

Figure 4. Representation of Q-Chat (A1–A10) variables.

Table 4. Chi-Square test for categorical variables.

Categorical Variables	*p* Value
Sex	0.0003848 ***
Ethnicity	1.834×10^{-6} ***
Jaundice	0.0224 ***
Family_mem_with_ASD	1
Relation	0.526
A1	2.2×10^{-16} ***
A2	2.2×10^{-16} ***
A3	2.2×10^{-16} ***
A4	2.2×10^{-16} ***
A5	2.2×10^{-16} ***

Table 4. *Cont.*

Categorical Variables	*p* Value
A6	2.2×10^{-16} ***
A7	2.2×10^{-16} ***
A8	2.2×10^{-16} ***
A9	0.89932
A10	0.73447

Note: *** denotes statistical significance.

Table 5 rejects the null hypothesis when the *p*-value is less than the significance level of 0.05. Therefore, the type of autism used has no impact on the Q-Chat score. Therefore, Q-Chat score columns were removed from the dataset for further analysis.

Table 5. ANOVA test for continuous variables.

Continuous Variables	*p* Value
Age	0.03165 ***
Q-Chat-score	0.34444

Note: *** denotes statistical significance.

Figure 5 presents a boxplot of the age attribute, which reveals the presence of outliers. To improve model accuracy, these outliers need to be removed before fitting the models. The "Age With Outliers" plot includes all data points, highlighting potential anomalies, whereas the "Age Without Outliers" plot provides a refined view by excluding these extreme values. This refinement allows for a clearer analysis of the typical age distribution within the dataset.
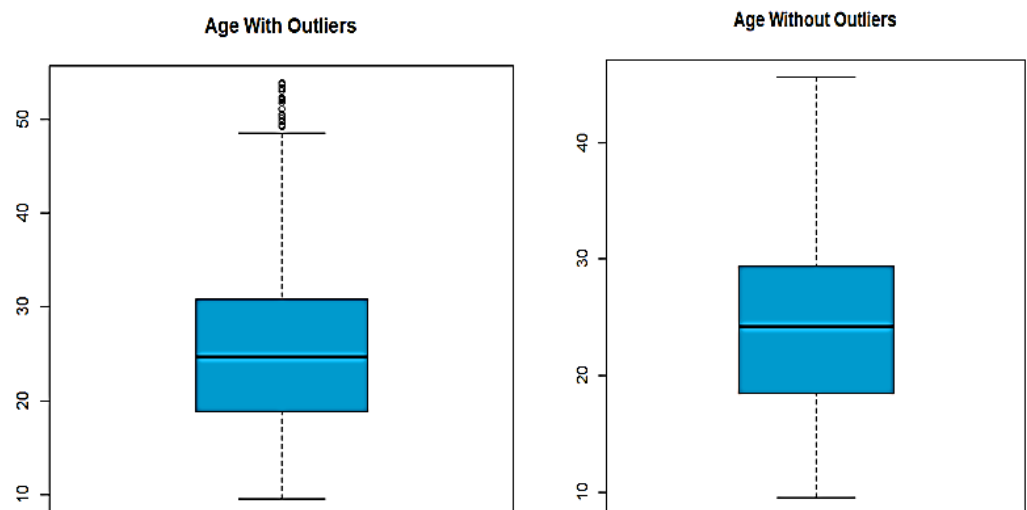

Figure 5. Identification and removal of outliers in the age variable using boxplot analysis.

The correlation plot in Figure 6 shows the relationship among the Q-Chat questions (A1–A10). Notably, variable A10 shows minimal correlation with the other variables, justifying its exclusion from further analysis. The plot highlights a positive correlation between most variables, with A1 displaying correlations of 0.46, 0.24, 0.25, 0.28, 0.37, 0.33, 0.21, 0.32, 0.13, and 0.61 with A2 through A10, respectively. However, the weak correlation of A10 with other variables supports its exclusion prior to model fitting.
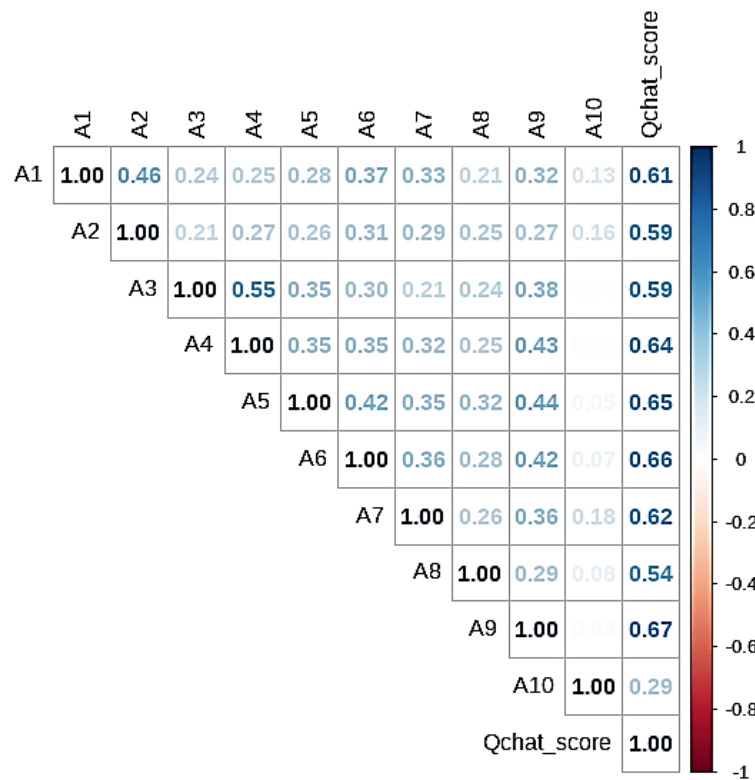
Figure 6. Correlation analysis of Q-Chat scores (A1 to A10) for autism-detection variables.

3.4. Data-Mining Techniques—Feature Selection

DM techniques encompass a diverse array of methods aimed at obtaining valuable insights from large datasets. These techniques include EDA, correlation analysis, anomaly detection through Boxplot, Chi-Square tests, and ANOVA. They are essential across various domains, from business intelligence to healthcare, enabling organizations to make data-driven decisions and extract actionable insights from complex datasets. The data-mining techniques applied in this study are summarized as follows.

3.4.1. Bivariate Analysis

After conducting a single-variable analysis, the subsequent stage involved a bivariate analysis to compare the two variables. This research strategy employs statistical methods to generate quantitative findings related to the dependent variable class, further supported by graphical representations. In this study, the following two types of bivariate analyses were performed:

1. Chi-Squared Test: This assessment examines categorical values to determine if a significant correlation exists between two categorical variables [34]. The Chi-Square test, a statistical method, evaluates the presence of a meaningful relationship between these variables [35]. The Chi-Square formula is as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \qquad (1)$$

Here, χ^2 denotes the Chi-Square statistic, O signifies the observed value, and E denotes the expected value.

Data for Chi-Square tests are typically presented in a cross-tabulation format, with each row representing a category of one variable and each column representing a category of another. It is essential that both variables originate from the same population and are categorical, such as class (Yes/No), sex (Male/Female), jaundice (Yes/No), ethnicity, and relation (Yes/No).

2. ANOVA: The Analysis of Variance (ANOVA) technique is used to evaluate mean differences between two groups for numerical variables. In this section, the ANOVA test was applied to continuous columns (age, Q-Chat square). The ANOVA test shows that the response variable varied according to the level of the categorical variable (or class). This hypothesis was as follows:

H$_0$: *The two variables are independent.*

H$_1$: *The two variables relate to each other.*

$$F = \frac{MST}{MSE} \tag{2}$$

where F is the ANOVA coefficient, MST is the mean sum of all squares owing to treatment, and MSE is the mean sum of squares owing to error.

3.4.2. Correlation Analysis

The correlation analysis explored the interrelationships among multiple statistical variables, quantifying the degree of linear association between pairs of variables [36]. This approach measures how closely related the variables are and whether they tend to change systematically in relation to one another. Correlation analysis is widely applied in fields such as finance, medicine, and social sciences to understand interdependencies between variables.

3.4.3. Outlier Detection

An outlier refers to a data point that significantly deviates from the majority of values in a dataset [37]. In the context of a boxplot analysis, an outlier is identified as any data point that falls outside the interquartile range. For this study, we examined the numerical variable "age" to assess its distribution, utilizing a boxplot to identify potential outliers.

3.5. Applied ML Methods

To improve ASD detection, various ML and DL algorithms [38–41] are applied to meticulously analyze data collected through mobile applications. The collected data were then fed into sophisticated algorithms designed to identify distinct patterns in children's behaviors. In this novel approach, we employed autism data that underwent an advanced preprocessing phase to enhance the dataset's representativeness. This sophisticated integration of mobile app technology and computational algorithms represents a comprehensive approach to understanding and analyzing children's behavioral dynamics.

3.5.1. LR with SHAP Analysis

LR, a statistical methodology, is adept at constructing regression models for response variables [42]. In logistic regression, $p(X)$ represents the probability that the response variable equals 1, given a set of predictor variables $X_1, X_2, \ldots, X_p$. This model estimates probabilities using the logistic function, ensuring that the output values always fall between 0 and 1, making it suitable for modeling binary outcomes. The logistic model is expressed as follows:

$$\log\left[\frac{p(X)}{1 - p(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{3}$$

Here

- X_j: Represents the jth predictor variable;
- β_j: Denotes the coefficient corresponding to the jth predictor variable.

On the right-hand side, the equation forecasts the log odds ratio of the response variable, assuming a value of 1. Hence, an LR model was established using the following equation:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}} \tag{4}$$

In this study, to explain the performance of the best-performing classifier, we used explainable AI through a method called SHAP, which interprets the output of ML models. SHAP is based on game theory and estimates the contribution of each feature in generating a model's output. After the classification process, a SHAP analysis was performed to identify the most essential features for achieving more accurate results. According to this analysis, A3 emerged as a key feature for producing the best outcomes with XGBoost [43]. The SHAP value formula is given by

$$\Phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{5}$$

where N denotes all feature sets and S represents a subset. Then, $S \cup \{i\}$ is the union of subset S of feature i. In this case, $v(S \cup \{i\})$ is trained on feature $S \cup \{i\}$, and $v(S)$ is trained with feature i left out.

3.5.2. Extreme Gradient Boosting

XGBoost, or Extreme Gradient Boosting, extends decision trees by incorporating multiple trees that work together to determine the final output, rather than relying on individual trees alone. XGBoost is a powerful ML algorithm known for its efficiency and accuracy in supervised learning tasks, particularly in classification and regression problems. It belongs to the ensemble learning family and is based on gradient boosting. The equation representing the objective function of XGBoost is

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where

- Obj is the overall objective function;
- n is the number of training instances;
- $L(y_i, \hat{y}_i)$ is the loss function that measures the difference between the actual target y_i and the predicted target $\hat{y}_i$;
- K is the number of weak learners (trees) in the ensemble;
- $\Omega(f_k)$ is the regularization term that penalizes complex models.

XGBoost iteratively adds new trees to minimize the objective function by using techniques such as gradient descent and exact or approximate algorithms for tree construction. This iterative process efficiently optimized the model for predictive accuracy.

3.5.3. Deep Models with Residual Connections and Ensemble (DMRCE)

Ensemble models are well-suited for ML because they combine the results of multiple models, offering a more robust prediction. While a single decision tree may provide a specific answer, a collection of trees or forests with different types of trees can deliver a more accurate and reliable result. In this project, we explore a combination of ensemble methods and DL by integrating a neural network architecture that incorporates residual networks (ResNet). ResNet introduces the concept of residual connections, or skip connections, which help to address the vanishing gradient problem in very deep networks. During the training of deep networks, error gradients tend to diminish as they propagate back through each layer, causing the gradients to approach zero in deeper layers. To counteract this, residual connections allow gradients to bypass layers with small gradients and propagate effectively

to subsequent layers. This helps maintain the flow of information and improves overall model performance. In this project, we aim to combine ensemble models with the ResNet approach to create a neural network that leverages both concepts, maximizing accuracy. The model architecture features a three-headed ensemble-type neural network followed by a deep dense layer with residual connections. This approach demonstrates how to build such models using TensorFlow and Keras.

In Figure 7, skip connections are shown as the fundamental mechanism behind residual networks. The skip connection links layer activations to subsequent layers by bypassing intermediate layers, thus forming a block structure. The omitted intermediate blocks are aggregated to generate residuals, enabling the network to focus on learning the residual mapping instead of each individual layer trying to learn the entire underlying mapping independently.

$$F(x) := H(x) - x \quad \text{which gives} \quad H(x) = F(x) + x \tag{7}$$

The derivative of the error with respect to x is expressed as

$$\frac{\delta E}{\delta x} = \frac{\delta E}{\delta y} \cdot \frac{\delta y}{\delta x} \tag{8}$$

$$= \frac{\delta E}{\delta y} \cdot (1 + F'(x)) = \frac{\delta E}{\delta y} \cdot \frac{\delta E}{\delta y} F'(x) \tag{9}$$
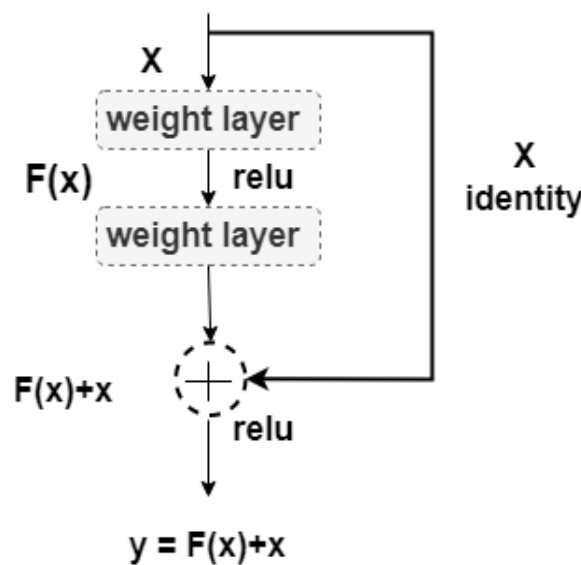

Figure 7. Residual networks.

3.5.4. FAST and Lightweight Automated ML (FLAML)

Automated ML (AutoML) refers to the process of automating various tasks involved in applying ML to real-world challenges. These tasks encompass everything from handling raw datasets to constructing ML models that are ready for deployment. In this study, AutoML is primarily employed to determine the most effective ML algorithm and corresponding parameters for the model. Using FLAML, this project automates common ML tasks from start to finish, with a variety of customization options. Furthermore, this study also performs comprehensive tuning of defined functions. The equation for AutoML can be expressed as

$$\text{AutoML} = \underset{\text{model} \in \text{models}}{\arg\min} \ \text{loss}(\text{model}(D_{\text{train}}), D_{\text{val}}) \tag{10}$$

where

- AutoML represents the automated ML process.
- Model denotes the ML model selected from a pool of potential models.
- Models refer to the set of potential models that can be considered during the AutoML process.
- D_{train} represents the training dataset.
- D_{val} represents the validation dataset.
- $\text{loss}(\cdot)$ represents the loss function used to evaluate the performance of the model on the validation dataset.

In this equation, the goal of AutoML is to determine the model (from a set of potential models) that minimizes the loss function when applied to the training data, while also being evaluated on the validation data. The model selection process includes hyperparameter tuning, feature selection, and engineering.

3.6. Hyperparameter Tuning

As part of this study's comprehensive research, we focused on fine-tuning the hyperparameters of various ML models. This objective was to identify sentence indicators and enhance the models' overall performance and generalization abilities. This process aimed to strike a balance between bias and variance, preventing overfitting. Ultimately, the goal was to determine the most precise and reliable hyperparameter settings for accurate sentence detection. In Table 6, we describe the model architecture and parameter settings for the different methods. To optimize model performance, we employed several strategies. Grid Search was used for a systematic evaluation of all possible hyperparameter combinations, while Random Search offered a quicker, randomized testing approach. FLAML was utilized for automated tuning within specific constraints, such _budget. Cross-validation was employed to ensure robustness, which is especially important in settings like logistic regression's split ratio. Additionally, Bayesian Optimization was used, which leverages probabilistic models to efficiently navigate complex parameter spaces. This technique proved particularly useful for intricate models, such as XGBoost. Together, these methods enhance model performance by optimizing parameter settings to achieve higher accuracy and generalizability.

Table 6. Parameter settings of different applied methods.

Classifier	Model Architecture and Parameters
LR	`family = binomial, split ratio = 80:20, list = FALSE, probabilities > 0.5`
FLAML	`estimator_list = XGBoost, log_file_name = autism.log, time_budget = 600`
DMRCE	`verbose = 1, min_lr = 0.00001, batch_size = 20, epochs = 100, activation = relu`
XGBoost	`learning_rate = 0.002, objective = binary:logistic, eval_metric = auc, max_depth = 10, alpha = 0.51, gamma = 1.92, reg_lambda = 11.40, colsample_bytree = 0.70, subsample = 0.83, min_child_weight = 2.55`

4. Results and Discussions

This section describes the results of the ML methods applied to autism detection using the proposed real-time framework.

4.1. Software and Hardware Configuration

The experimental setup used to develop and assess the applied ML and DL techniques is discussed here. Python 3.6 was used for building and evaluating the applied approaches. A network attack dataset was imported using the pandas module, and the models were

trained and tested with the scikit-learn module. For DL models, the TensorFlow API was employed. All experiments were conducted on Google Colab with a GPU backend, 13 GB of RAM, and 90 GB of disk space.

4.2. Results of Applied Evaluation Methods

The ML model determines the likelihood that each instance belongs to a specific class [44]. A confusion matrix was used to evaluate model performance, with other performance metrics derived from this matrix. Table 7 presents the confusion matrix for this autism-detection project, where TP represents true positives, FP indicates false positives, FN refers to False Negative, and TN denotes True Negative.

Table 7. Confusion matrix of predicted and actual autism.

Predict\Actual	Have Autism	Not Autism
ASD predicted	True Positive (TP)	False Negative (FN)
ASD not predicted	False Positive (FP)	True Negative (TN)

The bar charts in Figures 3 and 4 provide insights into the dataset, while the boxplot in Figure 5 reveals outliers in the continuous variable "age". ANOVA and Chi-Square bivariate analyses were conducted to identify significant variables for the model based on p-values. According to the Chi-Square test results in Table 4, variables such as family members with ASD, relation, A9, A10, and prior app usage were excluded from the dataset. Similarly, based on the ANOVA test in Table 5, Q-Chat score variables were also excluded. Figure 3 indicates that most individuals in the dataset have a higher likelihood of being diagnosed with autism, with a demographic predominantly from North America and composed largely of South Asian, Middle Eastern, and White European ethnicities. The gender distribution appears balanced, and healthcare professionals represent a significant portion of the dataset. Additionally, the variables A1 through A10, particularly at the "yes" response level, show a higher likelihood of autism. Based on the Chi-Square and ANOVA tests, variables A1, A2, A3, A4, A5, A6, A7, A8, age, sex, and ethnicity were found to be significant for the model.

Again, from the LR analysis in Table 8, the variables A1, A4, A5, A6, A7, A8, and jaundice (yes) show significant results based on the p-value with a 95% confidence interval. The odds ratio indicates the changes in the unit concerning the dependent variable. The coefficients indicate the beta coefficient estimates and their significance levels. For every one-unit change in A1, A2, A4, A6, A7, A8, age, and jaundice (yes), the log odds of autism decrease by -0.061, -0.037, -0.056, -0.081, -0.122, -0.034, -0.0006, and -0.058, respectively. In contrast, for a one-unit increase in A3, A5, and sex (male), the odds of autism increased by 0.043, 0.0887948, and 0.0105284, respectively.

Additionally, we observe significant and insignificant variables for predicting autism class (yes or no). A1, A4, A5, A6, A7, and jaundice (yes) yielded significant results, with p-values less than 0.05. In contrast, A2, A3, A8, age, and sex showed insignificant results.

Subsequently, in Table 9, the four ML models are compared based on the accuracy, precision, recall, and F1-score described in the appropriate evaluation with the ASD preprocessed dataset and confusion matrix based on correctly and incorrectly categorized dependent variables (classes).

In Table 9 and Figure 8, the results show that the DMLRS model outperformed other state-of-the-art methods, achieving 99% accuracy, 99% precision, 98% recall, and an impressive 97% F1-score, which is very close to 1, indicating it is a well-balanced model for this dataset. In comparison, the FLAML model achieved 94% accuracy, 94% precision, 93% recall, and a substantial 93% DMRCE, while the other models reported 88% accuracy, 64% precision, 73% recall, and a notable 68% F1-score. Finally, the XGBoost model provides 85% accuracy, 76% precision, 80% recall, and a substantial 78% F1-score.

Table 8. LR analysis to extract key predictive patterns.

Variable Name	Estimate	Odds Ratio (95% CI)	*p* Value	Relationship
Intercept	0.9593803	2.610079	2×10^{-16} ***	Significant
A1	−0.0611398	0.9406917	0.022261 *	Significant
A2	−0.0372514	0.9634339	0.166254	Insignificant
A3	0.0436941	1.044663	0.097515	Insignificant
A4	−0.0562393	0.9453129	0.033022 *	Significant
A5	0.0887948	1.092856	0.000686 ***	Significant
A6	−0.0815689	0.9216692	0.001979 **	Significant
A7	−0.1228839	0.8843663	1.33×10^{-5} ***	Significant
A8	−0.0343828	0.9662016	0.173490	Insignificant
Age	−0.0006173	0.9993829	0.322985	Insignificant
Sex (m)	0.0105284	1.010584	0.562028	Insignificant
Jaundice (yes)	−0.0584768	0.9432001	0.001709 **	Significant

Note: *, **, *** denotes statistical significance.

Table 9. Comparison of model performances.

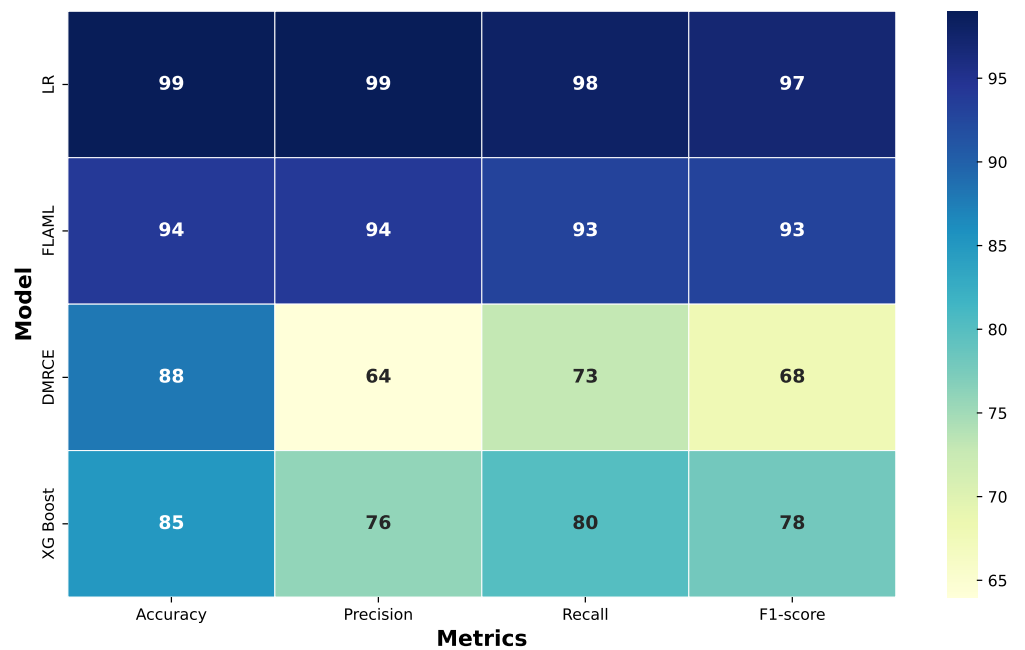| Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
DMLRS	99	99	98	97
FLAML	94	94	93	93
DMRCE	88	64	73	68
XGBoost	85	76	80	78

Figure 8. Heatmap for comparison of applied models.

Figure 9 shows a SHAP analysis of the important features used by the DMLRS technique. Based on the SHAP values of each feature in LR classifiers, the most significant

outcome include "A3_score", "A9_score", result, "A6_score" , and others. In addition, while most features improved the classification results of the classifiers, some features had a more substantial impact on the model's performance. SHAP analysis identifies key features influencing predictions, assisting clinicians in prioritizing factors that are crucial for patient care decisions.
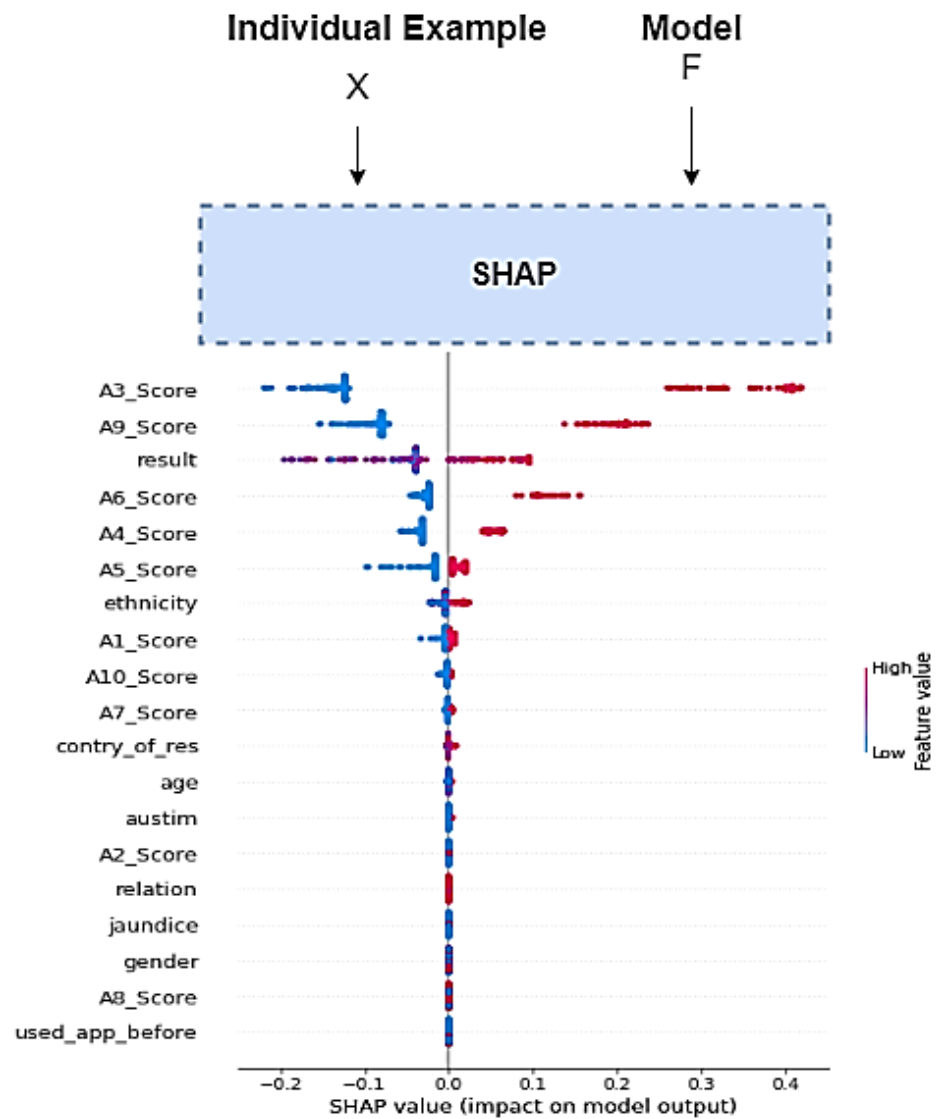

Figure 9. Unlocking DMLRS model's inner workings via SHAP.

Despite these findings, an LR approach is recommended, as it offers specific functions that can help predict autism. Combining multiple techniques will integrate all processes across the variables, allowing for a more comprehensive approach to predicting autism and enhancing the performance of various DM techniques in this context.

The confusion metrics for DMLRS, FLAML, DMRCE, and XGBoost are presented in Figure 10.

Again, in Figure 11, model accuracy and loss plot show for DMRCE as accuracy plot can be seen for the DL model of DMRCE. The accuracy plot shows the DMRCE model's learning progression over the training epochs. Rapid initial gains indicated fast learning, with a plateau suggesting convergence. The stability at high accuracy in later epochs reflects strong generalization, whereas fluctuations might indicate overfitting.
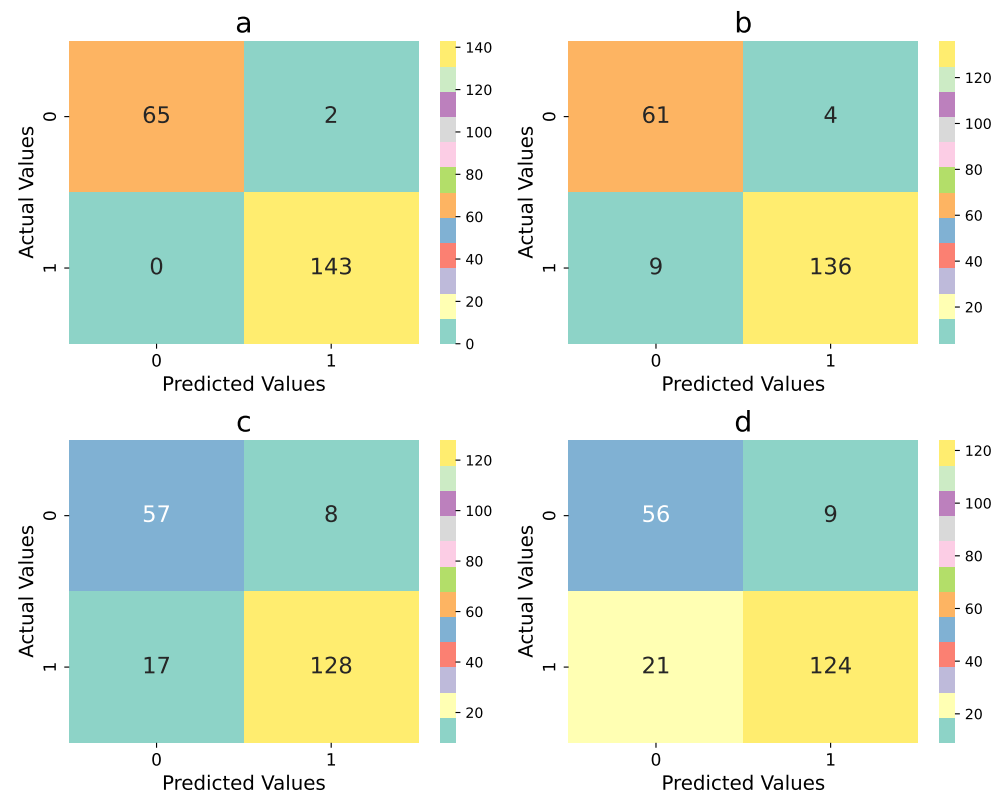
Figure 10. Confusion matrix of all applied models (**a**) DMLRS, (**b**) FLAML, (**c**) DMRCE, and (**d**) XGBoost.
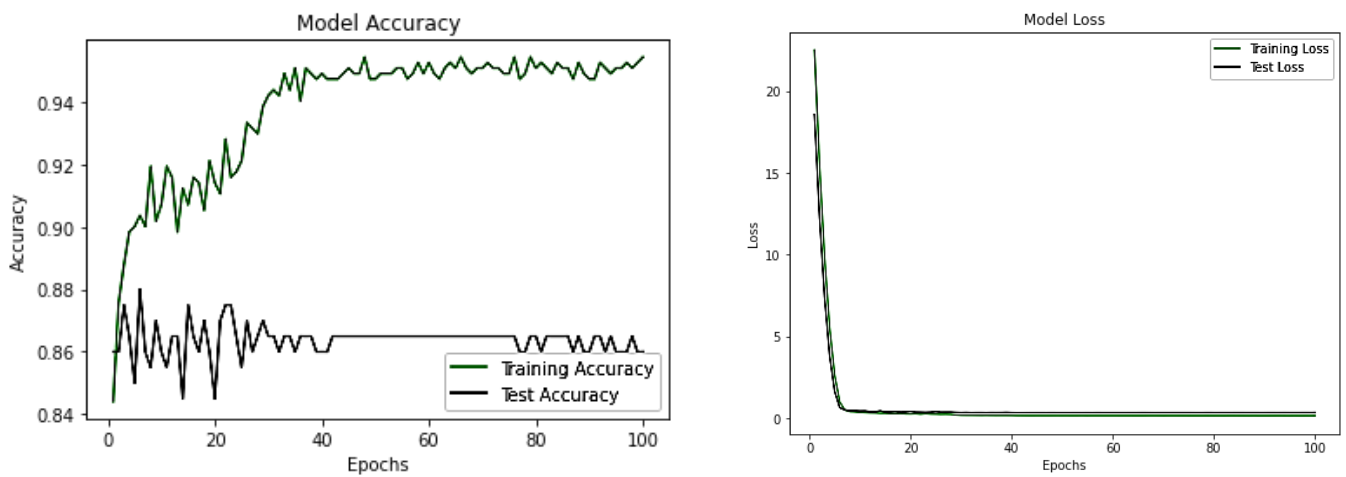

Figure 11. DMRCE model accuracy and loss to track the trends.

4.3. K-Fold Cross-Validation

K-fold cross-validation (K-fold CV) is crucial for assessing model generalization and robustness by providing a more reliable estimate of performance across different data splits, thus reducing the risk of overfitting compared with the single evaluation in the previous subsection. The k-fold CV process is mathematically represented as follows:

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} E\left(M_i, D_{\text{test}}^{(i)}\right) \tag{11}$$

where M_i is the model trained on all but the i-th fold, $D_{\text{test}}^{(i)}$ is the i-th fold used for validation, and $E(M_i)$ is the evaluation of model M_i on $D_{\text{test}}^{(i)}$. Table 10 displays the results of the 10-fold cv for each method. The DMLRS method has an impressive accuracy of 0.98 and a low standard deviation of 0.0049, indicating consistent performance across different subsets. The FLAML model showed a moderate accuracy of 0.91 with a standard deviation of 0.0042. The DMRCE and XGBoost models show accuracies of 0.85 and 0.89, with standard deviations of 0.0081 and 0.0037, respectively, indicating a reasonable level of consistency.

Table 10. CV of performance analysis.

Method	K-Fold	Accuracy	Standard Deviation
DMLRS	10	0.98	0.0049
FLAML	10	0.91	0.0042
DMRCE	10	0.85	0.89
XGBoost	10	0.0081	0.0037

4.4. Computational Complexity by Runtime

In Table 11, computational complexity is represented by the runtime(s) required to complete the computation. The DMLRS method is relatively fast, taking only 0.41 s. FLAML, DMRCE, and XGBoost were also efficient, with a runtime of 0.50, 0.65, and 0.62 s, respectively. The data suggest that there is a trade-off between accuracy and runtime and that the choice of method may depend on the specific requirements of the application in terms of speed and performance.

Table 11. Runtime computations.

Method	Runtime Computations (Seconds)
DMLRS	0.41
FLAML	0.50
DMRCE	0.65
XGBoost	0.62

4.5. Comparison with Previous Studies

To ensure a robust evaluation, we benchmarked the performance of our novel proposal against state-of-the-art techniques. This review is of a broad spectrum of cutting-edge methods developed over the past year. Notably, the performance scores of the various current approaches displayed differences, with the lowest accuracy recorded at 98.10%, indicating room for improvement. The proposed DMLRS approach stood out significantly, achieving a maximum accuracy of 99%. From Table 12, our proposed DMLRS work achieved a higher accuracy than the other papers. Therefore, the proposed model can be applied to other autistic prediction studies. Table 12 was reproduced by us following the same protocol (training and test data) but with a different dataset. We used the autism dataset from a secondary source and adhered to the same methodology to ensure the consistency and comparability of the results.

Table 12. State-of-the-art of existing research studies on ASD prediction.

Ref	Methods	Tools	Dataset	Accuracy	Sensitivity	Specificity
[43]	SVM, LG, Tree	Weka	29 attributes, 627 samples	97.70%	99%	94%
[45]	Naive Bayes, S.O.M., Neural Fuzzy, LVQ, Neural Network, K-means, Fuzzy C Mean	Developed	16 attributes, 100 samples	98%	95.26%	96.16%
[46]	SVM, LR, DT, Probabilistic variations	R, Weka	28 attributes, 4540 samples	97.27%	98%	89.39%
[47]	SVM, LR, DT	Scikit-Learn	65 attributes, 2925 samples	97.16%	97.22%	97.40%
[48]	SVM	Weka	65 attributes, 1726 samples	95.17%	87.95%	96.20%
[2]	NB, BG, CART, C4.5, KS, SVM, RT	Weka	4 datasets, 18, 23, 23, 23 attributes (1054, 509, 248, 1118 samples, respectively)	97.77%	97.66%	97.16%
[49]	Decision Tree, Random Forest	R	20 attributes, 1054 samples	91.74%	99%	92.39%
[50]	NB, SVM, KNN	Weka	20 attributes, 1054 samples	98%	92.39%	92.11%
Our work	LR, Auto ML, DMRCE, XGBoost	R, Python	20 attributes, 1054 samples	99%	98%	99%

4.6. Web-Based Autism Application System

Autism detection is a critical area of contemporary research. Traditional medical approaches to diagnosing autism can be prohibitively expensive, presenting challenges for the general population in accessing healthcare services. To address this issue, this study introduces a cost-effective and user-friendly solution: a web application for autism detection, developed using the Django framework, as illustrated in Figure 12. This web-based Autism Prediction System is designed to be accessible and easy to use. To utilize this service, individuals must register on the platform and create an account using their names, emails, and passwords. Once logged in, the system prompts users to input specific information aligned with autism dataset parameters to facilitate the prediction of autism. Upon submission of the required data, the web application analyzed the information and provided an immediate assessment. If indicators of autism are detected, the system displays the message "You have Autism"; otherwise, it confirms "You have no Autism". This intuitive interface simplified the autism screening process, making it accessible to a wider audience. Many users tested the app for usability, providing feedback to improve the interface and functionality for practical clinical use.

4.7. Limitations of Study

While this research introduces innovative ML methodologies for autism detection, it is important to acknowledge certain limitations. First, autism cannot be definitively diagnosed through the ten research questions (A1–A10) used in this application; thus, the system was designed to merely suggest the likelihood of autistic traits rather than provide a conclusive diagnosis. The reliance on self-reported data introduces potential biases, as the accuracy of the input directly affects the reliability of the predictions. Additionally, there is a need for more diverse data encompassing various regions and cultures to enhance prediction accuracy. The limited range of characteristics considered for ASD diagnosis in this model may not comprehensively cover all aspects of ASD, potentially affecting its efficacy. Finally, the current model's performance could be challenged as the sample size and number of variables increase, indicating a potential need for further refinement to maintain accuracy and reliability in more complex scenarios.

Figure 12. Newly developed Autism app screenshot and comparison of all models.

5. Conclusions and Future Directions

In the current era, the early detection of ASD presents a significant challenge in the fields of medical science, particularly in DM and ML. This study endeavored to construct a system capable of accurately predicting ASD. Initially, various DM, ML, and DL techniques were employed to identify pertinent variables for the model, with the aim of achieving the highest possible prediction accuracy. This study explored and evaluated four ML techniques combined with DL approaches and diverse statistical DM methods at different stages of research. The primary objective of this study was to accurately classify ASD cases (identifying whether an individual has ASD) to provide early intervention opportunities for individuals with ASD. The proposed DMLRS model obtained the highest accuracy of 99%, which not only demonstrates credibility but also highlights its significant efficiency. The proposed approach outperformed previous studies on autism detection, achieving remarkable accuracy and making a noteworthy contribution to the field. Additionally, a comparative analysis was conducted to identify the most reliable DM method for the model and web application system. In summary, this study proposes a model combining DM and ML techniques to analyze ASD datasets and facilitate early autism detection. The model utilized feature importance methods like ANOVA and Chi-Square to identify significant features. Subsequently, DMLRS, Auto ML, DMRCE, and XGBoost have been used for early-stage autism classification. Furthermore, the SHAP interpretation method was applied for the in-depth evaluation of DMLRS's outcomes. Notably, unlike most ASD datasets that are genetic, this study focused on behavioral variables, a unique approach that is not commonly used in existing ML research.

Future Work

In future research, scholars may leverage advanced data-mining techniques to further enhance the robustness and reliability of our model. The scope of this work can be expanded to address real-world challenges in autism by integrating additional data-mining models, streamlining the analysis process. Moreover, future studies should focus on expanding

datasets and gaining deeper insights into the characteristics associated with ASD, with an emphasis on utilizing primary data to improve outcomes.

Exploring the possibility of constructing hybrid classifiers by combining diverse methodologies is a promising approach. Moreover, enhancing the proposed framework could lead to the development of a more sophisticated system, benefiting both individuals and advanced healthcare systems for autism.

Author Contributions: Conceptualization, M.A.T.R., F.T.J., N.T., A.R., N.L.F., M.S. and S.W.L.; methodology, M.A.T.R., A.R., N.L.F., M.S. and S.W.L.; validation, F.T.J. and N.T.; formal analysis, M.A.T.R., A.R. and N.L.F.; investigation, F.T.J. and N.T.; data curation, M.A.T.R., F.T.J., N.T. and A.R.; writing—original draft preparation, M.A.T.R., F.T.J., N.T., A.R. and N.L.F.; writing—review and editing, M.S. and S.W.L.; visualization, M.A.T.R., A.R. and N.L.F.; supervision, M.S. and S.W.L.; funding acquisition, M.S. and S.W.L. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The dataset and source codes are available in Github repository at https://github.com/aintlab/Autism-Detection-Application (accessed on 16 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Won, H.; Mah, W.; Kim, E. Autism spectrum disorder causes, mechanisms, and treatments: Focus on neuronal synapses. *Front. Mol. Neurosci.* **2013**, *6*, 19. [CrossRef] [PubMed]
2. Uddin, M.J.; Ahamad, M.M.; Sarker, P.K.; Aktar, S.; Alotaibi, N.; Alyami, S.A.; Kabir, M.A.; Moni, M.A. An Integrated Statistical and Clinically Applicable Machine Learning Framework for the Detection of Autism Spectrum Disorder. *Computers* **2023**, *12*, 92. [CrossRef]
3. Autism Spectrum Disorder (ASD). Available online: https://www.nhsinform.scot/illnesses-and-conditions/brain-nerves-and-spinal-cord/autism-spectrum-disorder-asd (accessed on 2 February 2024).
4. Autism-Spectrum-Disorders. Available online: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (accessed on 2 February 2024).
5. Loftus, Y. Autism Statistics You Need to Know in 2022. 2022. Available online: https://www.autismparentingmagazine.com/autism-statistics/ (accessed on 2 February 2024).
6. Autism Spectrum Disorder. Available online: https://my.clevelandclinic.org/health/diseases/8855-autism (accessed on 2 February 2024).
7. Autism Spectrum Disorder (A.S.D). Available online: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd (accessed on 2 February 2024).
8. Wiggins, L.D.; Reynolds, A.; Rice, C.E.; Moody, E.J.; Bernal, P.; Blaskey, L.; Rosenberg, S.A.; Lee, L.-C.; Levy, S.E. Using Standardized Diagnostic Instruments to Classify Children with Autism in the Study to Explore Early Development. *J. Autism Dev. Disord.* **2015**, *45*, 1271–1280. [CrossRef] [PubMed]
9. Lord, C.; Rutter, M.; Le Couteur, A. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **1994**, *24*, 659–685. [CrossRef]
10. Abdelhamid, N.; Thabtah, F. Associative Classification Approaches: Review and Comparison. *J. Inf. Knowl. Manag.* **2014**, *13*, 1450027. [CrossRef]
11. Vos, T.; Allen, C.; Arora, M.; Barber, R.M.; Bhutta, Z.A.; Brown, A.; Carter, A.; Casey, D.C.; Charlson, F.J.; Chen, A.Z.; et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **2016**, *388*, 1545–1602. [CrossRef]
12. Comer, R.J.; Comer, J.S. Fundamentals of Abnormal Psychology, 9th ed. Available online: https://www.amazon.com/Fundamentals-Abnormal-Psychology-Ronald-Comer/dp/1319126693 (accessed on 16 April 2024).
13. Raj, S.; Masood, S. Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. *Procedia Comput. Sci.* **2020**, *167*, 994–1004. [CrossRef]
14. Thabtah, F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics Health Soc. Care* **2019**, *44*, 278–297. [CrossRef]
15. Abdullah, A.A.; Rijal, S.; Dash, S.R. Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD). *J. Phys. Conf. Ser.* **2019**, *1372*, 012052. [CrossRef]
16. Alteneiji, M.R.; Mohammed, L.; Usman, M. Autism Spectrum Disorder Diagnosis using Optimal Machine Learning Methods. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, e100. [CrossRef]

17. Tartarisco, G.; Cicceri, G.; Di Pietro, D.; Leonardi, E.; Aiello, S.; Marino, F.; Chiarotti, F.; Gagliano, A.; Arduino, G.M.; Apicella, F.; et al. Use of Machine Learning to Investigate the Quantitative Checklist for Autism in Toddlers (Q-CHAT) towards Early Autism Screening. *Diagnostics* **2021**, *11*, 574. [CrossRef] [PubMed]

18. Baranwal, A.; Vanitha, M. Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models. In Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 24–25 February 2020; IEEE: New York, NY, USA, 2020; pp. 1–7. [CrossRef]

19. Akter, T.; Satu, M.S.; Khan, M.I.; Ali, M.H.; Uddin, S.; Lio, P.; Quinn, J.M.W.; Moni, M.A. Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders. *IEEE Access* **2019**, *7*, 166509–166527. [CrossRef]

20. Musa, R.A.; Manaa, M.E.; Abdul-Majeed, G. Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children Using Data Mining Techniques. *J. Phys. Conf. Ser.* **2021**, *1804*, 012089. [CrossRef]

21. Shahamiri, S.R.; Thabtah, F. Autism AI: A New Autism Screening System Based on Artificial Intelligence. *Cogn. Comput.* **2020**, *12*, 766–777. [CrossRef]

22. Hossain, M.D.; Kabir, M.A.; Anwar, A.; Islam, M.Z. Detecting autism spectrum disorder using machine learning techniques. *Health Inf. Sci. Syst.* **2021**, *9*, 17. [CrossRef]

23. Vakadkar, K.; Purkayastha, D.; Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. *SN Comput. Sci.* **2021**, *2*, 386. [CrossRef]

24. Sadaf Hriti, N.; Shaer, K.; Nafis Momin, F.M.; Mahmud, H.; Kamrul Hasan, M. Autism Classification using Visual and Behavioral Data. *medRxiv* **2021**. [CrossRef]

25. Joudar, S.S.; Albahri, A.; Hamid, R.A.; Zahid, I.A.; Alqaysi, M.; Albahri, O.; Alamoodi, A. Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: A systematic review of current trends and open issues. *Artif. Intell. Rev.* **2023**, *56*, 53–117. [CrossRef]

26. Albahri, A.; Joudar, S.S.; Hamid, R.A.; Zahid, I.A.; Alqaysi, M.; Albahri, O.; Alamoodi, A.; Kou, G.; Sharaf, I.M. Explainable Artificial Intelligence Multimodal of Autism Triage Levels Using Fuzzy Approach-Based Multi-criteria Decision-Making and LIME. *Int. J. Fuzzy Syst.* **2024**, *26*, 274–303. [CrossRef]

27. Joudar, S.S.; Albahri, A.; Hamid, R.A. Intelligent triage method for early diagnosis autism spectrum disorder (ASD) based on integrated fuzzy multi-criteria decision-making methods. *Inform. Med. Unlocked* **2023**, *36*, 101131. [CrossRef]

28. Joudar, S.S.; Albahri, A.S.; Hamid, R.A. Triage and priority-based healthcare diagnosis using artificial intelligence for autism spectrum disorder and gene contribution: A systematic review. *Comput. Biol. Med.* **2022**, *146*, 105553. [CrossRef] [PubMed]

29. iamSam5, Laxman Naik, Suryansu Dash, Tensor Girl, and Vijayabharathi. ML Olympiad—Autism Prediction Challenge. 2022. Available online: https://www.kaggle.com/competitions/autism-prediction (accessed on 16 April 2024).

30. Thabtah, F.; Peebles, D. A new machine learning model based on induction of rules for autism detection. *Health Inform. J.* **2020**, *26*, 264–286. [CrossRef] [PubMed]

31. Autism Traits Detection System. 2021. Available online: https://play.google.com/store/apps/details?id=com.rezanet.intelligentasdscreener (accessed on 16 April 2024).

32. Rony, M.A.T.; Satu, M.S.; Whaiduzzaman, M. Mining significant features of diabetes through employing various classification methods. In Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 27–28 February 2021; IEEE: New York, NY, USA, 2021; pp. 240–244.

33. Hassan, M.M.; Rony, M.A.T.; Khan, M.A.R.; Hassan, M.M.; Yasmin, F.; Nag, A.; Zarin, T.H.; Bairagi, A.K.; Alshathri, S.; El-Shafai, W. Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness. *IEEE Access* **2023**, *11*, 132196–132222. [CrossRef]

34. Islam, M.S.; Hasan, A.J.; Rahman, M.S.; Yusuf, J.; Sajol, M.S.I.; Tumpa, F.A. Location agnostic source-free domain adaptive learning to predict solar power generation. In Proceedings of the 2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG), Wollongong, Australia, 3–6 December 2023; IEEE: New York, NY, USA, 2023; pp. 1–6.

35. Sajol, M.S.I.; Hasan, A.J. Benchmarking CNN and Cutting-Edge Transformer Models for Brain Tumor Classification Through Transfer Learning. In Proceedings of the 2024 IEEE 12th International Conference on Intelligent Systems (IS), Varna, Bulgaria, 29–31 August 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.

36. Gogtay, N.J.; Thatte, U.M. Principles of correlation analysis. *J. Assoc. Physicians India* **2017**, *65*, 78–81.

37. Hubert, M.; Vandervieren, E. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* **2008**, *52*, 5186–5201. [CrossRef]

38. Ishtiaq, A.; Munir, K.; Raza, A.; Samee, N.A.; Jamjoom, M.M.; Ullah, Z. Product Helpfulness Detection with Novel Transformer Based BERT Embedding and Class Probability Features. *IEEE Access* **2024**, *12*, 55905–55917. [CrossRef]

39. Abbas, M.A.; Munir, K.; Raza, A.; Samee, N.A.; Jamjoom, M.M.; Ullah, Z. Novel Transformer Based Contextualized Embedding and Probabilistic Features for Depression Detection from Social Media. *IEEE Access* **2024**, *12*, 54087–54100. [CrossRef]

40. Naseer, A.; Amjad, M.; Raza, A.; Munir, K.; Samee, N.A.; Alohali, M.A. A Novel Transfer Learning Approach for Detection of Pomegranates Growth Stages. *IEEE Access* **2024**, *12*, 27073–27087. [CrossRef]

41. Raza, A.; Munir, K.; Almutairi, M.S.; Sehar, R. Novel Transfer Learning Based Deep Features for Diagnosis of Down Syndrome in Children Using Facial Images. *IEEE Access* **2024**, *12*, 16386–16396. [CrossRef]

42. Sperandei, S. Understanding logistic regression analysis. *Biochem. Medica* **2014**, *24*, 12–18. [CrossRef]

43. Wall, D.P.; Kosmicki, J.; Deluca, T.F.; Harstad, E.; Fusaro, V.A. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl. Psychiatry* **2012**, *2*, e100. [CrossRef] [PubMed]

44. Wang, C.; Wu, Q.; Liu, X.; Quintanilla, L. Automated Machine Learning & Tuning with FLAML. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 14–18 August 2022; pp. 4828–4829. [CrossRef]

45. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]

46. Schopler, E. Toward Objective Classification of Childhood Autism: Childhood Autism Rating Scale (CARS). *J. Autism Dev. Disord.* **1980**, *10*, 91–103. [CrossRef] [PubMed]

47. Pancerz, K.; Derkacz, A. Consistency-Based Preprocessing for Classification of Data Coming from Evaluation Sheets of Subjects with ASDs. In Proceedings of the Position Paper Federated Conference on Computer Science and Information Systems, Lodz, Poland, 13–16 September 2015; Volume 6, pp. 63–67. [CrossRef]

48. Pedregosa, F.; Weiss, R.; Brucher, M. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

49. McNamara, B.; Lora, C.; Yang, D.; Flores, F.; Daly, P. Machine Learning Classification of Adults with Autism Spectrum Disorder. 2019. Available online: http://rstudio-pubs-static.s3.amazonaws.com/383049_1faa93345b324da6a1081506f371a8dd.html (accessed on 16 April 2024).

50. Bala, M.; Prova, A.A.; Ali, M.H. Prediction of Autism Spectrum Disorder Using Feature Selection and Machine Learning Algorithms. In *Proceedings of the International Conference on Computational Intelligence and Emerging Power System, Ajmer, India, 9–10 March 2021*; Springer: Singapore, 2022; pp. 133–148. [CrossRef]