AI POWERED AUTISM PREDICTION:

Submitted by:

|  |  |
| --- | --- |
| Areeha Zainab | Zukhruf Riaz |
| 4225-FBAS/BSSE/F21 | 4241-FBAS/BSSE/F21 |
| Department of software Engineering | Department of software Engineering |
| International Islamic University | International Islamic University |  |

## **Supervisor:** Dr. Umara Zahid

## umara.zahid@iiu.edu.pk Lecturer Department of Software Engineering International Islamic University

## ****Course Title:**** Semester Project

**Date of Submission:**  
June 2025

Abstract:  
Autism Spectrum Disorder (ASD) is a complicated neurodevelopmental disorder that frequently goes untreated due to limitations in existing screening methods. This study proposes a scalable, AI-powered autism screening system that utilizes behavioral, demographic, and clinical variables to aid in early-stage prediction. An integrated dataset was created by combining publically available autism screening sources and performing data cleaning, normalization, and feature selection utilizing information gain. Multiple classification models were trained using a total of 15 high-impact characteristics, including age, family history, and AQ-10 screening items.Among them, XGBoost performed the best, with an accuracy of 89%, precision of 77%, and recall of 85%. To increase inter-pretability, SHAP (SHapley Additive Explanations) was utilized to display feature-level contributions, demonstrating that screening scores, age, and family history were the most influential predictors. The built FastAPI-based online system offers a simple interface for real-time screening help. While not a replacement for clinical diagnosis, the technique provides early risk identification and may be especially valuable in impoverished areas. This study reveals how inter-pretable machine learning can improve access to autism screening and help caregivers and professionals make timely decisions.

## Keywords

Autism Spectrum Disorder (ASD); Machine Learning; Feature Selection; Information Gain; XGBoost; SHAP Explainability; Autism Screening; Questionnaire-based Prediction; AI in Healthcare; Interpretability.

# Introduction:

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental disorder that impairs social communication, behavior, and cognitive flexibility. Its symptoms commonly emerge in childhood and last into adulthood, with different severity and appearance. Despite rising awareness, ASD remains underdiagnosed or diagnosed late, particularly in areas with inadequate clinical resources. Traditional diagnostic approaches, such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R), are time-consuming, need skilled specialists, and are sometimes unavailable in low-resource environments. This highlights the crucial need for scalable, efficient, and easily available early screening techniques.

The overlap of symptoms with other neurodevelopmental or psychiatric problems hampers the early and correct diagnosis of ASD. Early diagnosis is crucial because it allows for interventions during critical developmental windows, which improves long-term outcomes. However, due to restricted access to diagnostic services and the subjective nature of behavioral assessments, many people, especially adults, go untreated. This study seeks to overcome these issues by using machine learning (ML) to create a low-cost, interpretable, and accessible ASD screening method.

Although several studies have used machine learning approaches to predict autism, the majority of them focus on pediatric datasets and do not provide explainable results or interactive tools. Furthermore, few systems combine behavioral screening tools such as the AQ-10 with demographic and clinical data in real time. This study fills that gap by presenting a unique AI-based screening system that combines the AQ-10 questionnaire with other variables such as age, family history of autism, and neonatal jaundice. The most important characteristics are picked and trained on various machine learning models using entropy-based information gain.

The major goal of this study is to create and test an interpretable autism prediction system based on the FastAPI framework, with XGBoost as the principal classifier. SHAP (SHapley Additive ExPlanations) is used to show feature contributions, making the model's judgments clear to both users and medical practitioners. The device is designed to act as an early-stage screening assistance, which is especially useful in impoverished areas where clinical examinations are sparse.

The rest of the paper is organized as follows: Section 2 provides a review of related literature, Section 3 describes the proposed methodology and system architecture, Section 4 details data preparation and feature engineering, Section 5 discusses model evaluation and interpretability, and Section 6 concludes the study with future research directions.

# Literary Review:

Recent breakthroughs in artificial intelligence (AI) and machine learning (ML) have sparked renewed interest in the creation of automated systems for healthcare screening, including Autism Spectrum Disorder (ASD). ASD is difficult to diagnose, and typical techniques necessitate professional competence, are time-consuming, and frequently unavailable in low-resource settings. As a result, academics have proposed a number of AI-based algorithms to improve autism prediction accuracy and scalability.

## AI in Autism Screening:

Several research have investigated the use of typical machine learning techniques for ASD identification. Khudhur et al. [1] tested machine learning classifiers such as Decision Tree, Random Forest, Naive Bayes, and Support Vector Machines (SVM) on age-segmented autism datasets. Their research indicated that Random Forest performed consistently well across all age groups. However, they also identified a significant flaw: the lack of interpretability in most models, which reduces trust and therapeutic utility.

Wang et al. [2] created a web-based diagnostic interface employing behavioral and demographic data. Although the system was very usable and automated, it lacked transparency in decision-making, limiting its application in real-world clinical situations. Al-Attar et al. [3] addressed issue by introducing rule-based reasoning into their hybrid AI system, which was combined with behavioral assessments. While their model achieved reasonable accuracy, it lacked scalability and interactive deployment, which limited its widespread use.

## Feature Selection and Interpretability in Health AI:

Recent research has underlined the importance of appropriate feature selection for improving model performance and reducing computing cost. Uddin et al. [4] developed SHAP (SHapley Additive ExPlanations) to evaluate the results of machine learning models for ASD detection. SHAP helps clinicians trust AI suggestions by explaining how individual factors contribute to a prediction. Their findings indicated increased clinician trust, however the system was only tested in controlled research settings, not in real-time applications.

Other studies on healthcare prediction have shown that information gain and entropy-based feature selection strategies are beneficial. For example, the authors of [5] showed that selecting high-impact features based on entropy significantly improved classification accuracy for conditions such as diabetes and Parkinson's disease. This strategy decreases noise and redundancy in the dataset, which improves model focus and interpretation.

## Application Gaps in Age Inclusivity and Real-time Tools:

Jang et al. [6] undertook a large-scale evaluation of over 2,800 autism-related research papers and discovered a notable age imbalance: 94% of studies focused on people under the age of 20. Only 21% looked at adult autism, indicating a significant research deficit in understanding and predicting ASD in older groups.  
  
Furthermore, most existing ASD prediction models are trained and tested using static datasets. Few systems have real-time prediction capabilities, and even fewer combine known behavioral screening tools like the AQ-10 with clinical and demographic information in a single, interactive interface [7][8].

## Literature Gap and Study Contribution:

The preceding review reveals various holes. First, while many models are accurate, they frequently lack interpretability and are unsuitable for real-time or public-facing applications. Second, there is little emphasis on adult populations in ASD research, which limits generalizability. Third, few systems combine verified screening tools like AQ-10 with modern feature selection approaches like information gain, and even fewer use SHAP to show feature impact in real time.  
  
This study tackles these shortcomings by offering a FastAPI-based AI system that incorporates behavioral, demographic, and clinical data obtained through information acquisition and analyzed using SHAP. The system is deployable, useable by non-experts, and focuses on explainability, making it a unique addition to the present landscape of ASD screening tools.

# Proposed Solution:

This paper offers an AI-powered autism screening system that uses behavioral, demographic, and clinical data to generate early predictions about Autism Spectrum Disorder (ASD). The system is built on the FastAPI framework and includes a web-based user interface, Information Gain feature selection, a trained XGBoost classifier, and SHAP interpretability. The process is divided into numerous clearly defined steps, as shown below.

## System overview:

The total system architecture is intended to handle user inputs in real time and offer a quick ASD risk estimate with visual explanations. Figure 1 depicts the architecture of the proposed system. It includes the following components:

**The User Interface Module** is a web-based questionnaire in which users respond to 10 AQ-10 screening questions as well as basic demographic information.  
  
**The Data Preprocessing Module** converts categorical replies into numerical feature vectors, normalizes the data, and maintains consistency with the training dataset.  
  
**The Feature Selection Module** uses Information Gain to keep the 15 most relevant features that contribute significantly to autism prediction.  
  
**The AI Prediction Module** uses a pre-trained XGBoost model to estimate the likelihood of ASD based on input features.  
  
**The Interpretability Module** uses SHAP to show feature contributions and give users with an explanation of the prediction.  
  
**Recommendation Module:** Based on the results, users are advised whether to seek expert advice or conduct further evaluation.

This modular design ensures that data moves systematically and efficiently through the system, from user input to ultimate output.

## Real-time Questionnaire Input:

The system incorporates a real-time questionnaire that collects user replies to the ten AQ-10 behavioral screening questions. Each question includes four response options: Definitely Agree, Slightly Agree, Slightly Disagree, and Definitely Disagree. These responses are converted to numerical scores that are consistent with the dataset's encoding. Additional information—such as age, gender, family history of autism, jaundice at birth, and the user's relationship to the individual—is collected concurrently.  
  
As soon as the user submits the form, the data is transformed into a structured feature vector and sent to the prediction module. This process is automated, takes only a few seconds, and provides fast feedback.

## Feature Selection using Information Gain:

Given the large number of available features, a feature selection technique was used to narrow down the model to the most useful inputs. Information Gain was estimated for each candidate feature in the dataset. This statistic measures how much knowing a feature reduces uncertainty in predicting ASD. The 15 traits with the greatest information gain were retained:  
  
AQ scores range from A1 to A10.  
  
**Clinical and demographic:** age, result, autism (family history), jaundice, relationship.  
  
This technique improves prediction accuracy while simultaneously reducing computing complexity and overfitting.

## AI Model – XGBoost Classifier:

At the heart of the system is an XGBoost classifier, a high-performance gradient boosting model noted for its durability and speed. This model was chosen following a comparison with several classifiers like Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, and Random Forest.  
  
The XGBoost model takes the selected 15 features as input and returns a binary classification: autistic or not autistic. A probability threshold is used to transform the output score to a discrete class label. The model was trained on 80% of the dataset, then verified on the remaining 20%. Grid search and cross-validation were used to tune hyperparameters for optimal performance.

## Model Interpretability with SHAP:

SHAP (SHapley Additive exPlanations) was added into the system to increase transparency and user trust. SHAP values quantify how each feature affects the model's prediction for a single user.  
  
Once the forecast is produced, a SHAP summary graphic is created to visually show which features had the most influence. In most situations, the AQ screening score, age, and family history are the primary contributors. This component fills the gap between raw prediction output and meaningful human interpretation, making the technology therapeutically useful and ethically accountable.

## Recommendation and Feedback Module:

Based on the prediction, the algorithm makes a personalized recommendation. If the risk of ASD is high, the user should seek professional advice or a formal diagnostic evaluation. For low-risk forecasts, the system may recommend additional monitoring or general developmental direction. These results are provided in a clear, non-alarming manner, with a disclaimer stating that the tool is a screening aid and not a replacement for professional diagnosis.

## Backend Infrastructure and Deployment:

The system runs on a FastAPI backend server, which performs both model inference and SHAP explanation creation. The frontend connects with the backend using HTTP POST requests, which send user inputs and get structured prediction results. To ensure consistency, all preprocessing stages (such as encoding and scaling) mirror the training pipeline.

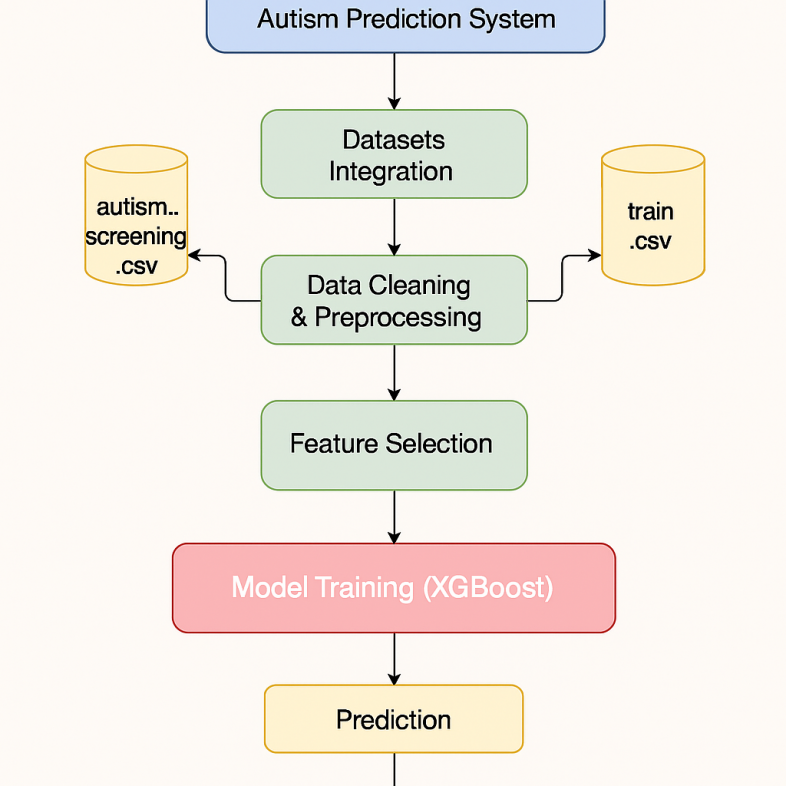
****

Figure 1: System Architecture

This proposed methodology results in a well-integrated and user-friendly AI system for early ASD screening. It stresses real-time usability, clinical relevance, interpretability, and scalability—all of which are critical for real-world deployment and impact.

# Dataset and Data Collection:

This study makes use of a consolidated dataset gathered from publicly available sources, with a specific focus on autism screening data. The main datasets used were

## Dataset Description:

Autism\_screening.csv and train.csv from Kaggle.  
  
The AQ-10 (Autism Spectrum Quotient) Questionnaire is a well-known 10-item screening instrument developed to identify features linked with Autism Spectrum Disorder (ASD).  
  
After carefully combining and cleansing various sources, we created a uniform dataset called autism. A csv file was prepared with 1,455 instances. Each instance represents an individual subject and is annotated with features such as behavioral scores (from AQ questions), demographic attributes (e.g., age, gender), clinical history (e.g., jaundice, family history of autism), and the final classification label indicating ASD likelihood.

## Data Source and Collection Method:

The data was gathered entirely from publicly available databases, with no private or personally identifiable information (PII) included. The AQ-10 questionnaire data was utilized to create the real-time questionnaire interface. Each user answer in the live system correlates to an AQ-10 item. The original datasets were collected from Kaggle using open licenses and were previously anonymised.  
  
In our suggested system, users input responses in real-time using a digital form fashioned after this questionnaire, which includes the following fields.

· 10 AQ-10 screening questions

· Age

· Gender

· Jaundice at birth (Yes/No)

· Family history of autism (Yes/No)

· Country of residence

· Whether they have used the screening app before

· Relationship to the individual being screened

To build the model, we aggregated all of the above data into a structured dataset, with each row representing one individual's complete response profile and screening result.

## Dataset Attributes:

The table below lists the main attributes used in the final dataset:

| **Feature** | **Description** |
| --- | --- |
| A1\_Score to A10\_Score | Responses to AQ-10 screening questions (scored 0 or 1) |
| Age | Age of the individual (in years) |
| Gender | Gender identity (encoded as binary) |
| Jaundice | Whether the subject had jaundice at birth (Yes = 1, No = 0) |
| Autism | Whether the subject has a family history of autism (Yes = 1, No = 0) |
| Country\_of\_res | Country of residence |
| Used\_app\_before | Whether they’ve used the app before (Yes = 1, No = 0) |
| Relation | Relation of the responder to the subject |
| Result | Total AQ score from 0–10 |
| Class/ASD | Target variable: 1 (ASD) or 0 (Non-ASD) |

Note: Non-informative factors such as ethnicity and age\_desc were eliminated during preprocessing due to their low association with the intended label.

## PreProcessing and Cleaning:

Several data cleaning steps were performed to ensure model readiness:  
**Invalid entries**, such as exaggerated age values, were eliminated.  
**Missing values** in categorical fields were either replaced or eliminated based on severity.  
**Label encoding** was applied to binary text fields, such as Yes/No to 1/0.  
**One-hot encoding** was used for multi-class categorical data such as related and gender.  
**Normalization** scaled numerical features (e.g., age, outcome) to a range of [0, 1].  
**Feature Selection** Using Information Gain ranked all features, and the best 15 were selected for training. These included the ten AQ questions, age, result score, jaundice, family history, and relationship.  
To ensure class balance, the dataset was divided into 80% training and 20% testing subsets using stratified sampling. The model was trained and verified using this split.

## Ethical and Privacy Considerations:

The dataset utilized in this study is public and anonymized, so no personally identifiable information was accessed or retained. The data are responses from volunteer, anonymous participants in previous research or surveys.  
  
In the live version of the screening tool, user data is not kept or transferred until explicit authorization is given. All inputs are analyzed locally for real-time prediction, and the system contains a disclaimer stating that it is a screening tool rather than a diagnostic solution.

To ensure ethical AI usage:

* The dataset was balanced to reduce bias
* Care was taken to avoid overfitting to specific populations
* The model’s decision logic is made visible by employing SHAP explanations.

## Data Summary:

The final processed dataset had the following characteristics:

* **Total instances**: 1,455
* **Positive ASD cases**: 697 (47.9%)
* **Negative ASD cases**: 758 (52.1%)
* **Number of final features used for training**: 15
* **Missing data after cleaning**: 0%

This dataset offered a solid foundation for developing a machine learning model with good predictive capability and clinical relevance.

# AI Model Selection and Training:

To overcome the classification challenge of predicting Autism Spectrum Disorder (ASD), we tested a variety of machine learning models, including Decision Tree, Random Forest, Naive Bayes, Logistic Regression, and XGBoost. Among these, XGBoost (Extreme Gradient Boosting) was chosen as the principal model because of its greater accuracy and generalization.  
  
With structured/tabular data, XGBoost is an effective ensemble technique based on gradient boosting that is renowned for managing intricate feature interactions, minimizing overfitting through regularization, and producing excellent prediction performance. Its ability to work with both categorical and numerical inputs made it ideal for our dataset, which contained AQ-10 scores, demographic information, and medical history.

## Training Procedure:

Following preprocessing, the dataset for this investigation includes 1,455 samples. The data was randomly divided into two subsets: 80% for training and 20% for testing, with class balance preserved in both. No extra validation sets were developed because the testing set was only used for final performance evaluation after training.

## Hyperparameter Training:

Given the time constraints and emphasis on showing feasibility, we used the default hyperparameters given by the XGBoost Python library without doing substantial hyperparameter customization. This enabled us to immediately evaluate the model's baseline performance on the selected features.

## Training Details:

Model training was carried out utilizing the XGBoost Python API. The training method used 80% of the data, while predictions were created from the remaining 20%. No cross-validation, early halting, or ensemble averaging techniques were used. Our goal was to determine whether a conventional implementation of XGBoost might produce clinically useful findings depending on the features we chose.

## Features Used:

As indicated in the Feature Selection section, we used 15 features chosen based on information gain. These included the AQ-10 behavioral screening scores (A1\_Score to A10\_Score), age, screening results, jaundice at birth, family history of autism, and the respondent's relationship to the person being evaluated. Using this subset reduced data dimensionality while keeping the most important predictive information.

## Evaluation During Training:

The model was validated on the testing dataset utilizing key classification metrics such as accuracy, precision, and recall. The XGBoost model obtained 89% accuracy, 77% precision, and 85% recall, making it the top-performing model among those tested. These metrics demonstrate the model's ability to effectively differentiate ASD and non-ASD instances using the given features.

## Special Techniques:

No advanced modeling techniques, such as transfer learning, ensemble averaging, or pipeline integration, were employed. The study focuses on creating and testing a single high-performing model with carefully chosen input features and a tried-and-true algorithm.

# Experiment and Evaluation:

## Experiment Design:

To assess the effectiveness of our autism prediction system, we trained and tested several machine learning models on a curated dataset consisting of autism screening questionnaire responses and demographic information. To ensure reproducibility, the dataset was split into 80% training and 20% testing using train\_test\_split and a fixed random seed (42). The test set, which was unseen during training, was utilized to evaluate model generalization and prediction performance.  
  
Six classification models were tested, including K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, XGBoost, and Logistic Regression. The project sought to determine the best accurate and balanced model for identifying suspected autism patients.

## Evaluation Metrics:

We assessed all models using three main categorization metrics:  
  
**Accuracy** is the percentage of correct predictions across all test instances.  
  
**Precision** is the proportion of genuine positives among all positive predictions, indicating false alarm control.  
  
**Recall** that the fraction of true positives found among all actual positives is critical in reducing missed cases in autism identification.  
  
Given the healthcare context, recall is especially important to guarantee that fewer probable ASD cases are missed, while precision remains relevant to decrease undue concern.

## Results Presentation:

All models were trained with the same 15 manually picked characteristics. Initially, Information Gain directed feature selection, but based on domain relevance and diagnostic value, the final feature set included all ten AQ scores, age, outcome, autism, jaundice, and connection.

**Table 1: Performance of Classification Models on Test Set**

| **Model** | **Accuracy** | **Precision** | **Recall** | **Features** |
| --- | --- | --- | --- | --- |
| KNN | 0.8522 | 0.7143 | 0.7237 | 15 |
| Naive Bayes | 0.8660 | 0.7079 | 0.8289 | 15 |
| Decision Tree | 0.8625 | 0.7143 | 0.7895 | 15 |
| Random Forest | 0.8935 | 0.7711 | 0.8421 | 15 |
| **XGBoost** | **0.8969** | **0.7738** | **0.8553** | 15 |
| Logistic Regression | 0.8832 | 0.7625 | 0.8026 | 15 |

The XGBoost model outperformed all other models, with the highest accuracy (89.69%) and the best balance of precision (77.38%) and recall (85.53%). These findings imply that XGBoost's capacity to handle complicated feature interactions and regularization enabled it to generalize more effectively than simpler models. Random Forest followed closely behind, offering it a strong alternative despite somewhat lower ratings.  
  
On the other hand, Naive Bayes achieved the highest recall (82.89%) following XGBoost but at the expense of lesser accuracy, indicating a proclivity to over-predict positive cases — this may be acceptable in early screening where erroneous positives are less critical than false negatives.

## System Interface and User Interactions:

To improve usability and accessibility, the trained AI model was deployed using an interactive web-based interface built with FastAPI. This interface allows users—such as doctors, researchers, or caregivers—to enter responses to the AQ-10 questionnaire as well as additional demographic and historical data such as age, gender, family history of autism, and more.  
  
When the form is submitted, the system examines the input using the trained XGBoost classifier and immediately shows a prediction result showing whether the individual is "At Risk of ASD" or "Not At Risk." Following Figure shows how this result is displayed in a clean, user-friendly report screen.

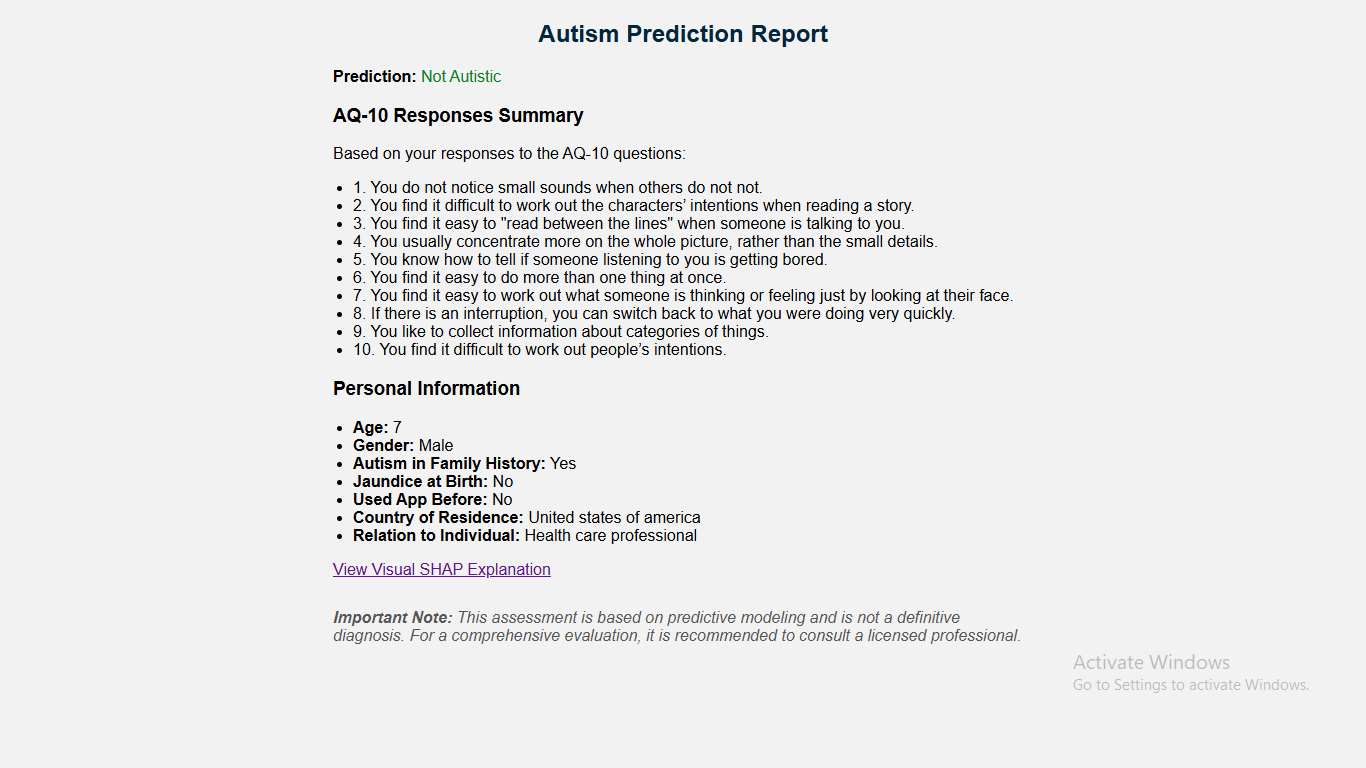


Figure 2: Prediction Screen

To improve model transparency and interpretability, the system generates a SHAP-based visual explanation. SHAP (SHapley Additive Explanations) identifies the most important features that contribute to the model's prediction for a given example. A high A9\_Score or a low age value, for example, may considerably shift the forecast to "At Risk". The SHAP plot, included in the interface and saved as an interactive HTML file, allowing users to investigate individual feature contributions. Figure 3 shows a sample SHAP visualization.

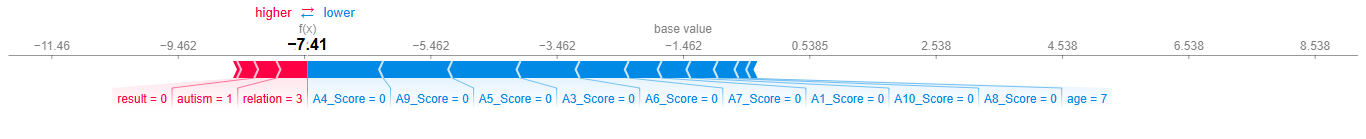


Figure 3: SHAP Display of Important Features

The prediction report and SHAP explanation work together to provide customers with not only a classification but also an understanding of the underlying thinking, enabling increased trust and interpretability in a sensitive environment such as autism screening.

## Interpretation of Results:

The selected features, particularly the AQ scores and key demographic factors, had a significant impact on model performance. The trade-off between precision and recall was evident across all models, but the XGBoost classifier produced the best balanced and robust results. No sophisticated approaches, such as cross-validation or ensemble averaging, were used, which may provide future opportunities for improvement.  
  
XGBoost was chosen as the final deployment model and saved to joblib for incorporation into the real-time screening system.

# Limitations:

While the results of our autism prediction method utilizing machine learning are encouraging, many limitations must be mentioned in order to present a balanced viewpoint and define the extent of our findings.

## Data Limitations:

One major restriction is the dataset's extent and representativeness. Although the dataset was compiled from publicly available sources, it may not fully represent the global population's variety. The majority of samples originate from specific demographic regions, and particular groups, such as women or people from underrepresented nations, may not be well represented. Furthermore, the dataset was relatively tiny, which may limit the model's ability to generalize to unknown real-world instances. The findings presented here should thus be viewed as preliminary and subject to further validation using larger and more diverse datasets.

## Feature and Scope Limitations:

Although Information Gain influenced feature selection, we discovered that some features relevant in real-world clinical autism diagnosis were given lower importance. To solve this, a hybrid method was taken, combining top-ranked features with manual selection based on domain relevance. Nonetheless, the current model is limited to 15 characteristics, which are primarily generated from self-reported AQ-10 questionnaire scores and a few demographic data. The model does not capture all potential ASD signs since it excludes biometric or behavioral data such as genetic markers, speech/language patterns, and brain imaging.

## Methodological Limitations:

Due to time constraints, no advanced cross-validation techniques were used, and the model was tested using a single train-test split. This could cause some variation in the reported performance, as results may fluctuate slightly depending on the split used. Furthermore, while XGBoost has excellent predictive accuracy, it functions as a black-box model, which may limit its interpretability—particularly in healthcare settings where explainability is critical for confidence and clinical adoption.

## System Limitations:

The system is currently implemented as a local prototype and has not yet been tested for real-time load or multi-user deployment. Predictions are currently generated on a single backend, with no integration into a scalable cloud environment. In practical circumstances, particularly where several concurrent users are expected (e.g., public health platforms), the system may experience performance bottlenecks or latency concerns unless improved or redeployed in a distributed environment.

## User Feedback and Validation:

This study did not include a rigorous user trial or clinical validation among healthcare professionals. The system has not been evaluated in a live situation with actual users or caregivers, limiting our ability to evaluate its usability, interpretability, and practical impact. Real-world feedback and clinical collaboration are required to confirm its efficacy as a screening tool.

## Generalizability:

The model has been trained particularly for binary classification of Autism Spectrum Disorder using chosen questionnaire and demographic data. Its ability to execute related tasks, such as identifying ASD subgroups, early intervention outcomes, and broader neurodevelopmental problems, has yet to be tested. Applying the same model architecture to different prediction tasks would necessitate retraining and further experiments.

## Mitigation and Transparency:

We acknowledge that the constraints listed above influence how the results should be viewed. For example, the reported accuracy of 89.69% using XGBoost is promising, but it may differ when evaluated on more diverse or noisy real-world data. While the hybrid feature selection strategy increases relevance, it also introduces subjectivity. These issues emphasize the need for additional validation, model development, and larger testing in future study.

# Conclusion:

This study aimed to create a machine learning-based autism screening system using publicly available datasets and self-reported questionnaire data. The primary goal was to create an accurate and accessible tool that could aid in the early detection of Autism Spectrum Disorder (ASD) risk, particularly in settings where clinical resources are limited.  
  
We found 15 significant attributes that contributed considerably to ASD prediction using a combination of automated and manual feature selection methods, including AQ-10 questionnaire results, age, and family history. We constructed and assessed six different machine learning models and discovered that the XGBoost classifier performed the best, with an accuracy of 89.69%, precision of 77.38%, and recall of 85.53%. These findings indicate the efficacy of integrating intelligent feature selection with robust classification approaches in screening scenarios.

The hybrid feature selection approach is one of the work's most significant contributions. While Information Gain was initially utilized to rank qualities, manual changes were made to favor variables with known real-world relevance to ASD, thus integrating data-driven insight with domain understanding. Furthermore, this study provides a ready-to-use, deployable machine learning model that may be included into digital screening tools or health apps.  
  
The findings indicate that a lightweight, questionnaire-based approach can achieve excellent predicted accuracy without requiring substantial clinical data or costly tests. This suggests that, if verified further, such a model could be a useful preliminary screening tool, particularly in under-resourced healthcare settings or rural locations where professional diagnosis may be delayed.

Looking ahead, various extensions are possible. To boost generalizability, future study may consider collecting a larger and more demographically varied sample. Cross-validation approaches would improve result dependability, while support for SHAP or LIME explanations could improve forecast transparency, which is a necessary condition for clinical use. Finally, conducting real-world user trials with caregivers or healthcare workers would provide vital feedback and help to validate the system's utility.  
  
Finally, our study shows that a machine learning model trained on carefully selected variables can accurately predict autism risk using questionnaire data, indicating a possible path for scalable and accessible early screening options.

# Acknowledgements:

The authors would like to express his gratitude to Dr. Umara Zahid for her essential assistance, feedback, and encouragement throughout this research endeavor. Their insights helped shape the methodology and final results of this study.  
  
We are also grateful to the International Islamic University for providing the academic resources and environment required to perform this research. In addition, the authors thank peers and classmates for their useful talks during the early phases of growth.  
  
This work was conducted as part of an academic requirement and did not receive any outside support.

# References:

[1] J. Jang, J. L. Matson, H. L. Adams, M. J. Konst, P. E. Cervantes, and R. L. Goldin, "What are the ages of persons studied in autism research: A 20-year review," Research in Autism Spectrum Disorders, vol. 8, no. 12, pp. 1756–1760, Dec. 2014. [Online]. Available: [https://doi.org/10.1016/j.rasd.2014.08.008](https://doi.org/10.1016/j.rasd.2014.08.008" \t "_new)

[2] D. D. Khudhur and S. D. Khudhur, "The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups," Measurement: Sensors, vol. 27, p. 100774, 2023. [Online]. Available: [https://doi.org/10.1016/j.measen.2023.100774](https://doi.org/10.1016/j.measen.2023.100774" \t "_new)

[3] R. Al-Attar, A. Fakhri, M. Ghalib, A. Alahmari, and S. Alharbi, "Autism AI: A new autism screening system based on artificial intelligence," Cognitive Computation, vol. 12, pp. 817–829, 2020. [Online]. Available: [https://doi.org/10.1007/s12559-020-09743-3](https://doi.org/10.1007/s12559-020-09743-3" \t "_new)

[4] M. Z. Uddin, F. Ali, and T. Sultana, "Early detection of ASD using explainable AI," Journal of Neuroscience Methods, 2024. [Online]. Available: [https://doi.org/10.1016/j.jneumeth.2024.110260](https://doi.org/10.1016/j.jneumeth.2024.110260" \t "_new)

[5] Y. Wang, D. Zhi, and Y. Qian, "A practical AI tool for diagnosing ASD," JMIR Medical Informatics, vol. 8, no. 5, p. e15767, May 2020. [Online]. Available: [https://doi.org/10.2196/15767](https://doi.org/10.2196/15767" \t "_new)

[6] U. Adnaan, "Autism screening dataset," Kaggle. [Online]. Available: [https://www.kaggle.com/datasets/umeradnaan/autism-screening](https://www.kaggle.com/datasets/umeradnaan/autism-screening" \t "_new) (accessed Jun. 20, 2025).

[7] Embrace Autism, "AQ-10: Autism Spectrum Quotient (10 items)," [Online]. Available: [https://embrace-autism.com/aq-10/](https://embrace-autism.com/aq-10/" \t "_new) (accessed Jun. 20, 2025).