Using Feature Selection Techniques to Investigate the Myth of Autism Spectrum Disorder

Albert Zheng*, He Zhu, Xinyi Hu, and Lan Yang

***Abstract*—autism spectrum disorder (ASD) is a developmental disorder that affects many people, especially children, with problems in communication and social life. Although many factors such as inherited gene mutation and environment influences may play important roles in autism, the actual causes of autism remain as myth. Without proper analysis for ASD, many people cannot get early detection of ASD and the public cannot fully understand ASD. Our goal is, through scientific investigation, to increase the public awareness on autism, so that better societal support could be provided to individuals with ASD traits. In this research, we analyze a children's autism screening dataset, apply feature selection techniques to identify key characteristics associated with ASD traits, and validate our findings with machine learning predictions. Our research revealed social responsiveness factors closely tied to ASD traits, and a strong link between autism assessment questionnaires and ASD traits.**

***Index Terms*—Autism spectrum disorder, categorical data, feature selection, correlation, chi square test, mutual information**

I. Introduction

Autism is spreading and affecting our future world. According to the Centers for Disease Control and Prevention, 1 in 100 children is diagnosed with an autism spectrum disorder, and about 1 in 44 children in the United States have an autism spectrum disorder (ASD). [1] ASD is a developmental disorder with the causes still as unsolved puzzle pieces. People with ASD often have problems with social communication and interaction, as well as restricted or repetitive behaviors or interests. The characteristics and degrees of autism symptoms vary greatly and differ case by case. Symptoms of autism traits can be subtle. Signs of autism in children often get ignored as they could be treated as delayed growth. However, it is crucial to notice the early symptoms of autism so that children exhibiting certain behaviors can receive special accommodations and societal support for learning, growth, and well-being. In this paper, we explore the ASD traits with data analysis and probe the important causes and symptoms with feature selection techniques. We hope, through this research endeavor, to gain better knowledge on autism traits, especially those in children, and to identify important characteristics of autism for early detection and increased awareness of autism.

While a great amount of research has been devoted to the detection of ASD, most of the work is laid in the fields of neuroscience and behavioral studies, such as trying to find DNA genes to unlock the mystery of autism [2, 3], improving autism screening and measurement methods [4, 5]. With the advancement of data science, data analytics-based studies are rising, such as empirical studies on autism intervention programs [5], using data mining tools to predict autism [6]. These research studies are aimed at better accuracy and/or early detection, however their probing is based on data collected via their own proprietary clinic studies or experimental processes.

Currently, the popular detection methods are questionnaires based. For example, Q-CHAT is a popular ASD assessment method for children. It is conducted by giving questionnaires to parents or caregivers. Such questionnaires-based assessment approaches can lead to several issues. For example, literacy and language barriers of parents or caretaker may cause misunderstanding of the questionnaires, and caregivers may not have full knowledge and heedful observation of child development. It is important to catch special features protruding autism prone children from daily life interaction.

Since it is difficult, but critical to differentiate children's behaviors between autism symptoms and delayed developmental skills, our studies attempt to extract the important features associated with ASD traits by analyzing an autism dataset acquired from We first conduct the exploratory analysis on the autism screening data, then apply feature selection techniques (correlation, feature of importance, Chi-Square test, and mutual information gain) to select a set of essential traits from all input variables of the dataset, i.e. perform dimensionality reduction. Finally, we validate our feature selection results by comparing the prediction results of ASD traits before and after dimensionality reduction. As autism symptoms are subtle, we try to sort out key characteristics of ASD traits to gain better knowledge on autism traits, increase global awareness on autism, and contribute some clues to the big puzzle of autism myth.

II. Exploratory Data Analysis of Autism Spectrum Disorder (ASD)

A. Overview of the Autism Screening Data

We obtained the dataset from kaggle.com, which was collected by an autism research group at the University of Arkansas [7]. The data collected includes assessments to children with autism signs by rating innate and social factors as well as Q-CHAT questionnaires. Q-CHAT is a quantitative assessment for children who may show some ASD traits [8]. The dimension of the dataset is (1985, 28), with each row representing a screening case, columns represent input

Manuscript received January 23, 2023; revised March 23, 2023; accepted May 5, 2023.

Albert Zheng is with the Academy of the Canyons, Santa Clarita, CA, USA.

He Zhu is with the Rancho Cucamonga High School, Rancho Cucamonga, CA, USA.

Xinyi Hu is with the Los Osos High School, Rancho Cucamonga, CA, USA.

Lan Yang, is with the California State Polytechnic Univ., Pomona, CA, USA.

*Correspondence: anpei.zheng@gmail.com (A.Z.)

doi: 10.18178/ijfcc.2023.12.4.609

variables. The input variables can be classified into two categories, variables of numerical values and categorical variables. There are 15 variables of numerical values, including ten Q-CHAT questionnaire scores represented as A1 to A10 respectively with 0 or 1 value, the sum of ten Q-CHAT scores, a social responsiveness scale (a score in range of 0 to 10 with 10 as the highest), age (from 0 to 17), and the childhood autism rating scale, i.e., CARS, rated on a seven-point scale (1, 1.5, 2.4) ranging from "within normal limits for that age" – coded as 0 to "severely abnormal for that age" – coded as 4, shown in Table I. The categorical variables include speech/language disorder, learning disorder, genetic disorder, depression, global developmental disorders, social/behavioral issues, anxiety, sex, ethnicity, jaundice at birth, family history of autism, who completed the test, and the ASD traits diagnosis, as shown in Table II. Most categorical variables have Yes/No two values except sex, ethnicity, and test takers.

We choose this dataset because it provides a large amount of data on children from age 0-17 who were screened for autism, which could be used to study various aspects of autism diagnosis and screening. The dataset may also be useful for developing and testing new screening tools or interventions for autism. Based on the information provided in the dataset, it appears to be relatively recent, as the latest entries in the dataset are from 2020. Table I shows the basic statistics of numerical data, where we observed: (1) the ten Q-CHAT test scores (A1 to A10, each Ai test receives a score of 0 or 1) and the QCHAT-Score (the sum of all Ai scores) are very consistent from the viewpoint of mean and standard deviation, which motivates us to further investigate whether these ten questions are correlated or independent, and whether we could identify a few key questions as important factors of ASD traits to simply the screening process; and (2) the average social responsiveness scale (a score from 0 to 10 with 10 as the highest) is very low, about 3.0, that indicates there may not be sufficient support or awareness towards autism from the society.

TABLE I: NUMERICAL SUMMARY OF CHILDREN'S AUTISM SCREENING DATA I

	Case #	A1	A2	A3	A4	A5	A6	A7
Count	1985	1985	1985	1985	1985	1985	1985	1985
Mean	993	0.30	0.24	0.21	0.27	0.28	0.31	0.35
Std	573.2	0.46	0.43	0.41	0.45	0.45	0.36	0.48

TABLE II: NUMERICAL SUMMARY OF CHILDREN'S AUTISM SCREENING DATA II

	Case#	A8	A9	A10	QCHAT Score	SR Scale	Age	CA RS
Count	1985	1985	1985	1985	1946	1976	1985	1985
Mean	993	0.24	0.26	0.45	4.23	3.07	9.62	1.70
Std	573.2	0.43	0.44	0.50	2.90	3.68	4.30	1.02

B. Visualization of Autism Screening Data

Table II shows the summary information of categorical variables. From Table II, we observed that most cases (suspected of ASD traits) involve all kinds of disorders, male dominate the group, there is no obvious family history linkage, and the screening tests are mostly completed by healthcare professionals (see Table III−Table IV).

TABLE III. PROFILE INFORMATION OF CATEGORICAL VARIABLES IN CHILDREN'S AUTISM SCREENING DATA I

	Speech disorder	Learn disorder	Genetic disorder	Depr.	GD disability	SB issues	Anxiety disorder
Count	1985	1985	1985	1984	1985	1971	1985
Unique	2	2	2	2	2	2	2
Top	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Freq	1057	1075	1013	1054	1054	1040	1051

TABLE IV: PROFILE INFORMATION OF CATEGORICAL VARIABLES IN CHILDREN'S AUTISM SCREENING DATA II

Sex	Ethnicity	Jaundice	Family history	Test takers	ASD Traits
1985	1985	1985	1985	1985	1985
2	16	2	2	6	2
M	White Europe	Yes	No	HP	Yes
1447	549	1536	1330	1233	1074

We explore more details with data visualization. Fig. 1 illustrates the differences in terms of age and jaundice (at birth). It shows that age 14 (when children turn to young adults) is the dominating age group when ASD traits are most concerned (a). This may indicate that a good number of cases were undetected until children were in early adolescence. Also, majority of those surveyed as well as identified with ASD traits were born with Jaundice (b)

We further study the society and environmental factors that affect ASD. Fig. 2(a) illustrates details of social responsiveness ratings where we observed that a large number of cases received no social responsiveness (i.e., scored 0), which is very alarming. Fig. 2(b) displays the ethnicity distribution of autism cases in which White European and Asian are two groups affected by autism significantly.

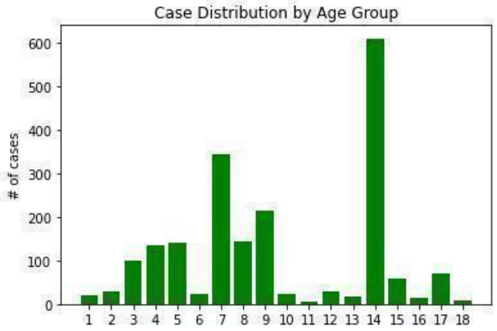


Fig. 1(a)

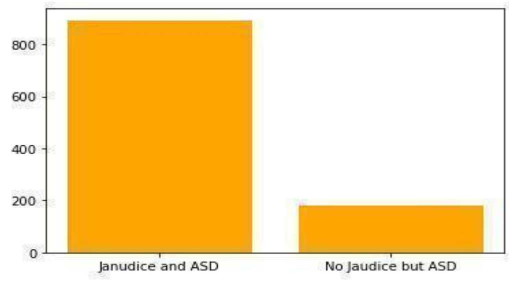
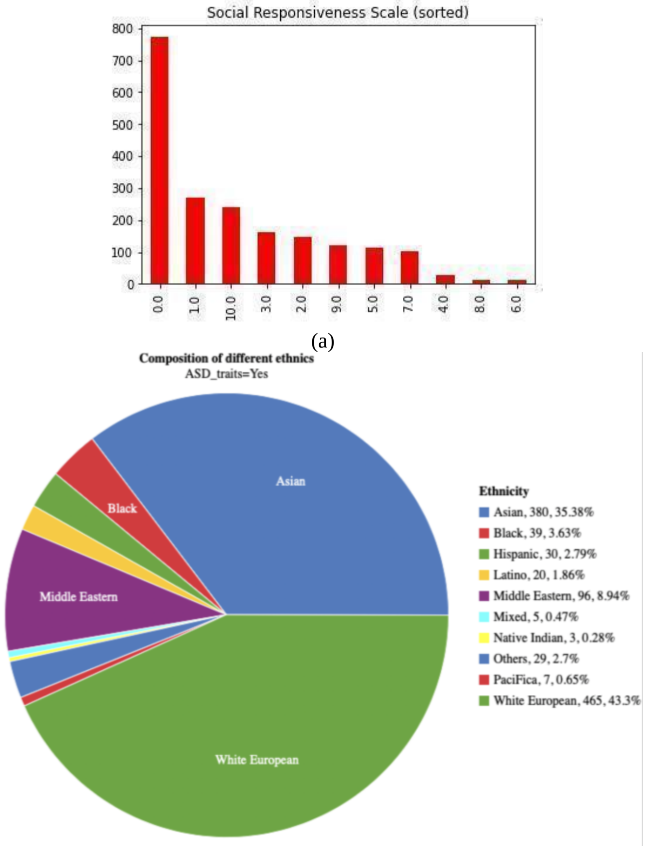


Fig. 1(b).

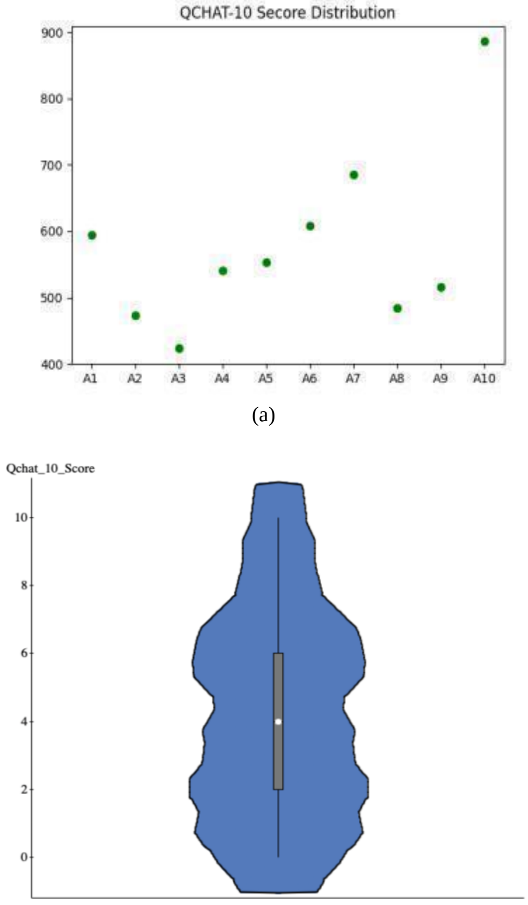
Fig 1. Relationship between personal factors and ASD traits. (a) Case distribution by age group. (b) Ethnicity distribution of autism cases.



(a)

(b)

Fig. 2. Relationship between social factors and ASD traits. (a) Social Responsiveness Rating. (b) Composition of Different Ethnics.



(a)

(b)

Fig. 3. Q-CHAT SCREENING results and distribution. (a) Q-CHAT screening results. (b) Q-CHAT screening distribution. detailed

Since ASD traits are mainly determined by Q-CHAT assessment, we study the relationship between the Q-CHAT questionnaires, A1 to A10 and ASD traits. Fig. 3(a) shows the screening results for each of the Q-CHAT questionnaires as well as the statistical summary (b) which reveals that the overwhelming high for A10, A1, A6, and A7 are all outstanding, while A3 is the lowest.

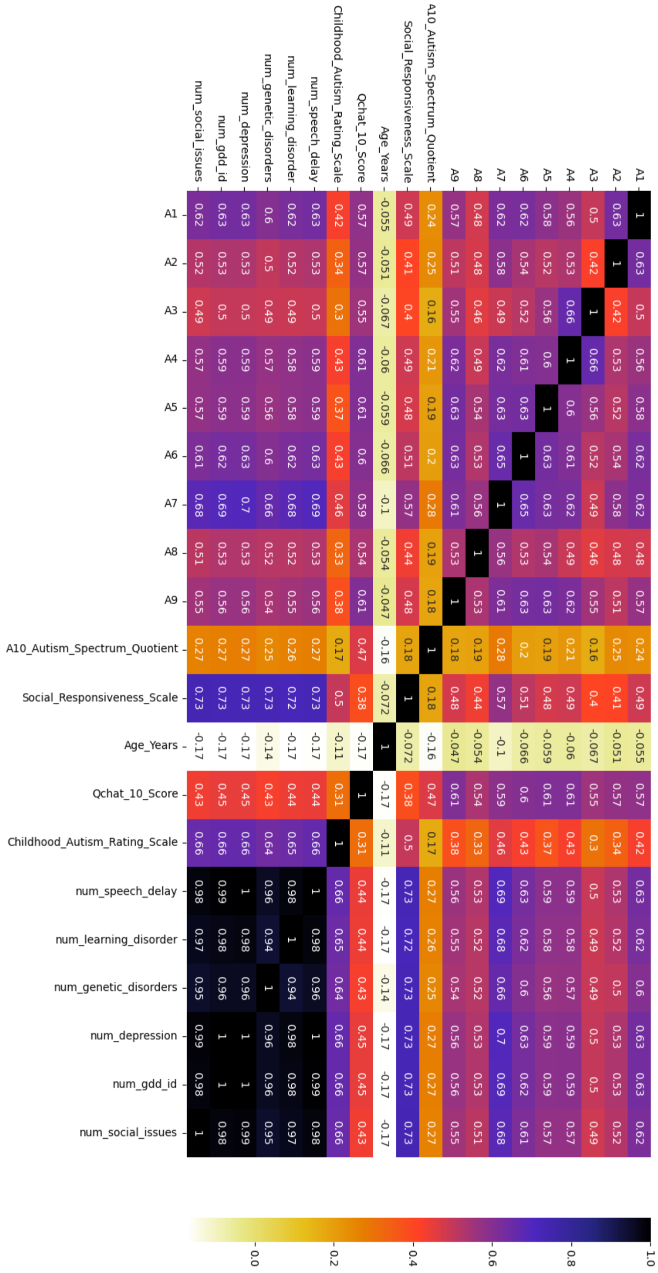


Fig. 4. Heatmap of correlation.

While our exploratory analysis revealed some useful information about the relationship between ASD traits and personal, societal, and screening factors, we would like to dig in-depth to extract the most important features from the 28 input variables. Hence, we apply the feature selection techniques to select a subset of key input variables (i.e., apply dimensionality reduction) from input variables that most relate to the ASD traits. We verify the feature selection using random forest prediction models.

III. Feature Selection For Predictive Modeling Of ASD Screening Data

A. Overview of Feature Selection Techniques

To reveal in-depth features, in this research we apply machine learning based feature selection techniques. Feature selection is the process to select the most relevant data, non-redundant input variables to use in developing machine learning predictive models. In constructing machine learning models, especially for high dimensional datasets, feature selection methods are first applied to reduce the input variables, and consequently decrease over-fitting, reduce training time, and improve accuracy. It can also facilitate a better understanding of the learning model or data, which is the primary goal of this research.

There have been several research publications on feature selection in machine learning. Cai et. al. surveyed supervised, unsupervised, and semi-supervised feature selection methods [9]. Hall addressed correlation-based feature selection methods for machine learning. Suresh *et al.* [10] applied feature selection techniques to detect autonomic Dysreflexia [11]. In this research, since most of our data are categorical, we use four methods, correlation, feature of importance, Chi-Square test, and mutual information statistic to explore feature selection [12].

B. Correlations Analysis

Correlation is a statistical measure that reveals whether the two variables are linearly associated or not. If two variables have a high rate of correlation, they could be considered as redundant variables. Fig. 4 shows the correlation heatmap calculated using Pearson's correlation coefficients [13]. We can observe from Fig. 4 that A4 and A5 are highly correlated, thus one of them could be redundant. However, Pearson's correlation coefficient has limitations. It does not work well for categorical variables that have more than two values nor for variables with non-linear relationships. Thus, the correlation measure only provides us limited information.

From the correlation analysis, we observed the following phenomena:

(1) Social/behavioral issues, depression, language disorders, and learning disorders are highly correlated (~99%). In Section III (B), we also observed from Figure 3.1 social responsiveness scale that the majority ranked the scale as 0, i.e. the lowest social responsiveness scale. This reveals that autism is a serious problem among those disadvantaged people. Social awareness is crucial to help children with autism.

(2) The Q-CHAT questionnaires are moderately related among A1 to A9. This means that the questions may be redundant, or a smaller set of questions may be used to set alert on autism. However, from correlation analysis, we are not able to tell specifics. The following sections will further address this observation.

(3) The Q-CHAT questionnaires A1 to A10 have low to very low correlation with Sex, Ethnicity, and family history. This indirectly shows that the Q-CHAT test is not biased.

C. Feature of Importance

Feature Importance calculates a score for all input variables, i.e., features, in a given machine learning model. The scores represent the "importance" of each feature. A higher score means that the specific feature will have a larger impact on the model in predicting an output variable. Feature of importance reveals the relationship between features (i.e., input variables) and the target (i.e., output variable) as well as the redundant features for the model. Different machine learning models may use distinctive ways of calculating feature of importance. In this research, the random forest classifier is used for feature of importance [14].

TABLE V: Top 5 and Bottom 5 Features of Importance When All variables Are Used for ASD_Traits Prediction

Variable	Top 5 features				
	Ethnicity	Sex	A7	A5	A6
Factor	0.224	0.135	0.079	0.068	0.062
Variable	Bottom 5 features				
	Depression	Anxiety disorder	Genetic Disorder	Global develop Delay	Jaundice
Factor	0.010	0.010	0.008	0.006	0.003

TABLE VI: Features of Importance When Only Ten Q-CHAT Test Questions Are Used for Predicting ASD_traits. Top 4 Factors in Bold

Variable	A1	A2	A3	A4	A5
factor	0.088	0.061	0.033	0.086	**0.138**
A6	**A7**	A8	**A9**	A10	**A6**
0.180	**0.140**	0.064	**0.135**	0.075	**0.180**

Feature of importance in a random forest is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. Table III shows the top 5 and bottom 5 importance factors when all 28 variables are considered, and Table IV is the result when only the ten Q-CHAT test questionnaires are used to predict ASD traits, and Table V displays the important social factors (i.e., all variables except the ten Q-CHAT test questionnaires are considered).

TABLE VII: Features of Importance when Only Social Factors (Non Q-CHAT Questions) Are Used for Predicting ASD_Traits

Var.	Age	Rating Scale	Lang. Disorder	Learning disorder	Global development. delay	Depression
factor	0.188	0.045	0.022	0.011	0.014	0.030
Genetic Disorder	Behavioral issues	Anxiety disorder	Sex	Ethnicity	Jaundice	Family history
0.032	0.171	0.026	0.180	0.329	0.005	0.095

Observations: (1) Ethnicity, Sex, and behavioral issues play important roles; (2) A5, A6, A7, A9 are important Q-CHAT questionnaires, which indirect proved our earlier interruption from the correlation analysis that the ten Q-

CHAT questionnaires may be redundant.

D. Chi Square Test

Chi-square (χ^2) test is a statistical model that provides numerical measurement to decide whether there is a relationship between two variables. It is effective in determining whether a categorical output variable (Y) is related or associated with another categorical input variable (X), which fits our dataset as it consists of mostly categorical variables including the output variable ASDTraits.

The Chi-square formula, as presented in [15], is shown as follows:

$$\chi_c{}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

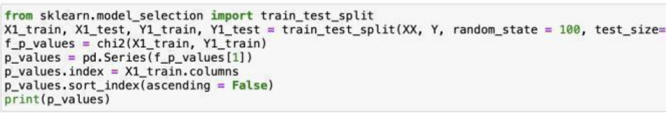
where:

C = Degrees of freedom
O = Observed values(s)
E = Expected values(s)

We implement the Chi-Square test using Python scikit-learn machine learning module, where we get the three values as output: the p-value, the statistical value, and the degree of freedom. A p-value is a statistical measurement used to validate a hypothesis against observed data. The lower the p-value, the greater the observed difference, i.e., the outcome is not closely related with the predictors. Typically, if $p <= 0.05$ the test hypothesis is false or should be rejected. In this paper, we demonstrated the Chi-Square measurement using two of the hypotheses that we put forward, H0 and H1, as illustration.

H0: Whether the age and ASD traits are independent.

H1: Whether Jaundice and ASD traits are independent. The code for the Chi-Square test is given in Fig. 5 and the program execution results are shown in Table 6.



```
from sklearn.model_selection import train_test_split
X1_train, X1_test, Y1_train, Y1_test = train_test_split(XX, Y, random_state = 100, test_size=
f_p_values = chi2(X1_train, Y1_train)
p_values = pd.Series(f_p_values[1])
p_values.index = X1_train.columns
p_values.sort_index(ascending = False)
print(p_values)
```

Fig. 5. Primary Code Segment for Chi-Square test.

TABLE VIII: P-VALUES FOR TWO HYPOTHESES

Hypothesis: Variable	p_value
H0: Age in Years	0.111208 (11.1%)
H1: Jaundice	0.012514 (1.25%)

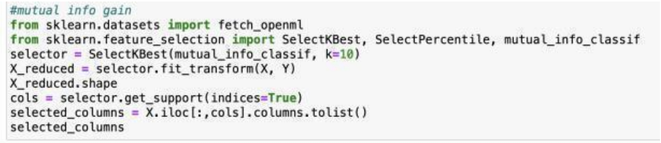
Conclusion: From Table VI, it is evident that there is an association between Jaundice and ASD traits at 5% significance level (p = 1.25 < 5%, i.e., hypothesis is false, which implies Jaundice and ASD traits are dependent) while Age is independent of ASD traits (p = 11.1% > 5%, thus the hypothesis is true.) Although in Section II (B) our exploratory analysis indicates that the majority identified with ASD traits at age 14, with the Chi-square test result, we are more confident about our speculation that this is caused by the delayed examination for ASD traits. Therefore, our studies set forth the importance of early ASD screening for children to display certain characteristics. We have identified a subset of features in Section III (C) and we further study that in Section III (E).

E. Mutual Information

Mutual information is a measure of statistical independence of variables. More powerful than the correlation analysis, mutual information can measure any type of relationship between variables, and not just linear associations. Thus, it is more suitable for our dataset as it contains both continuous and discrete variables. Mutual information (MI) is defined as the relative entropy between the joint distribution of the two variables and the product of their marginal distributions. A higher MI value indicates a large reduction in uncertainty, i.e., two variables are highly related. A zero value for MI means that the two variables are completely independent. Statistically, MI is defined as:

$$I(X,Y) = \sum_{X} \sum_{Y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)})$$

where I (X, Y) is the MI between variables x and y, the joint probability of the two variables is p (x, y), and their marginal probabilities are p(x) and p(y). [16] We used the univariate feature selection methods implemented in scikit-learn [8, 11] to select the top N features and the top P% features with the mutual information statistics.



```
#mutual info gain
from sklearn.datasets import fetch_openml
from sklearn.feature_selection import SelectKBest, SelectPercentile, mutual_info_classif
selector = SelectKBest(mutual_info_classif, k=10)
X_reduced = selector.fit_transform(X, Y)
X_reduced.shape
cols = selector.get_support(indices=True)
selected_columns = X.iloc[:,cols].columns.tolist()
selected_columns
```

Fig. 6. Primary code segment for calculating mutual information.

Results and analysis:

- X_reduced.shape = 10, i.e. the input dimension reduced from 28 columns to 10 columns, which cuts off 65% of original variables.
- Select top 10 result: selected_columns = ['A1', 'A2', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'S', 'Et']
- Analysis: Q-CHAT questionnaires (other than A3, A10) are important as well as Sex and Ethnicity. It is consistent with the feature of importance analysis by selecting Sex and Ethnicity as two most important categorical variables. For QCHAT questionnaires, however, the mutual information displayed an upset than that we obtained the feature of importance analysis, i.e. ['A1', 'A2', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9'] (from mutual information) vs. [‘A5’, ‘A6’, ‘A7’, ‘A9’] (from feature of importance).
- Select top 25%, Result: ['A1', 'A2', 'A4', 'A5', 'A6', 'A9']. This differs slightly from the feature of importance analysis, which yields [‘A5’, ‘A6’, ‘A7’, ‘A9’]. We suspect this is due to the high correlation among the Q-CHAT questionnaires that adds more uncertainty in mutual information analysis. In Section IV, we will further investigate the differences.

IV. PREDICTION OF ASD TRAITS AND VALIDATION OF FEATURE SELECTION

To validate the outcomes from the feature selection process, we have built up a machine learning prediction system based on a random forest model, with the dataset split into 70% training data and 30% validation data, to predict the ASD traits and analyze the following five scenarios.

- S1: predict ASDTraits based on all factors from input data file.
- S2: predict ASDTraits based on selected set of features ['A1', 'A2', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'S', 'Ethnicity'] that obtained from top 10 results of mutual information gain.
- S3: predict ASDTraits solely based on Q-CHAT screening test questionnaires i.e. A1 to A10 scores.
- S4: predict ASDTraits based on the subset of feature of importance [‘A5’, ‘A6’, ‘A7’, ‘A9’], considering the A1 to A10 Q-CHAT questionnaires only.
- S5: predict ASDTraits based on the selected subset of features ['A1', 'A2', 'A4', 'A5', 'A6', 'A9'] generated by the top 25% of mutual information gain.

Table VII shows the summary of prediction accuracy and confusion matrix from the above five scenarios. The results suggest that Scenario 1 predicts ASD Traits based on all the factors from the input data file which have a high accuracy of 86.24%, which is obviously true. However, Scenario 2 shows promising results, with a high accuracy of 83.22%, which is very competitive to Scenario 1. This concludes that the factors we obtained from feature of importance and mutual information gain reduced dimensionality from 28 to 10, i.e., used only 35.7% of original variables to receive competitive results. Scenario 3, 4, and 5 predicts ASD traits only based on QCHAT-10 questions. Scenario 3 (predict ASD traits based on A1 to A10 of Q-CHAT assessment) has better accuracy than any of the subset selected, i.e. Scenario 4 (predict ASD traits based on the subset of questionnaires generated by feature selection [‘A5’, ‘A6’, ‘A7’, ‘A9’]) and the top 25% from feature selection Scenario 5 (predict ASD traits based on the subset of questionnaires generated by feature selection ['A1', 'A2', 'A4', 'A5', 'A6', 'A9']), which indicates ethnicity, sex, and behavioral/genetic issues should also be considered in determining ASD traits. Scenario 4 and Scenario 5 have similar accuracy, in which we interpret that the subset selected via feature of importance is compatible with that selected with mutual information gain. The reason for variation of subset selection is largely due to the correlation among A1 – A10 questionnaires as revealed in Section III (B), which verifies the redundancy in Q-CHAT screening questionnaires.

TABLE IX: Prediction of ASD Traits and Validation of Feature Selection

Scenarios	Confusion Matrix				Accuracy (%)	Precision (%)	Recall (%)	FI Score (%)
	TN	FP	FN	TP				
S1: all factors	222	39	39	296	86.24	88.36	88.36	88.36
S2: top 25	174	87	12	323	83.22	78.78	96.42	86.71
S3: A1 to A10	257	4	109	226	81.01	94.16	67.46	78.60
S4: feature of importance Ai	250	11	132	203	76.01	94.86	60.60	73.95
S5: top 25%	250	11	119	216	78.12	95.15	64.48	76.87

It is important to not generate false information that leads to ASD traits not being identified and intervened in treatment at an early age. Therefore, we want to reduce the number of FNs as much as possible. The result for Scenarios 3, 4, and 5 from all the positive classes, the percentage for how many we predicted correctly is lower than Scenarios 1 and 2. Measures that take this into account are FN and Recall. Comparing the value of Scenarios 1 and 2, to Scenario 3, 4, and 5, FN is uncommonly higher. We can interpret that Q-CHAT is still the most important and effective method but including the categorical variable such as sex and ethnicity can greatly enhance the value of Recall. Sex and ethnicity for uncertainty estimation to improve prediction accuracy.

V. Conclusion and Future Work

In this study, we investigated the mystery of autism Spectrum Disorder by using feature selection techniques. Specifically, we applied three different feature selection methods to a dataset containing various behavioral and demographic features of individuals with and without ASD. We used feature selection methods to validate the robustness of our results. We focused on both demographic and behavioral features, which enabled us to investigate the role of various factors in ASD. Our study highlighted the importance of gender and developmental disorders, which could inform future research and clinical practice.

Overall, our study has important implications for understanding and diagnosing ASD. By identifying the most important features associated with ASD, we can better understand the underlying mechanisms and develop more accurate and effective diagnostic tools. Additionally, our findings can help clinicians and educators to identify individuals at risk for ASD and provide appropriate support and interventions.

There is a continued increase in the prevalence of autism spectrum disorder (ASD) worldwide. With greater public awareness on autism, it can help not only those individuals with autism to lead independent lives, but also enrich our society by accepting the wonderful differences that we all have. Our research takes on a data analysis approach to study the important features of ASD traits. Our studies discovered the following linkages and impacts on autism children:

- Innate factors: male children as well as children born with jaundice are more prone to ASD traits.
- Social factors: a lot of autism cases were ignored until children to adolescent ages and social responsiveness to autistic children is very low. Children with ASD traits also exhibited many other syndromes such as depression, social/behavioral issues, and developmental disorders.
- The ten Q-CHAT screening questionnaires are moderately correlated, and a subset of four Q-CHAT questions can perform early detection of ASD traits effectively.

Our future work will start with a survey on the autism community on how music and art activities could help children with ASD traits to stimulate cognitive functioning and then plan for local/community events by using music and art activities for remediation of speech/language and social skills.

Conflict of Interest

The authors declare no conflict of interests.

Author Contributions

He Zhu and Albert Zheng designed data analysis models. Xinyu Hu and He Zhu interpreted the analysis results. Albert Zheng implemented data analysis. All authors participated in paper writing and approved the final version.

References

[1] CDC. Data and Statistics on Autism Spectrum Disorder. [Online]. Available: https://cdc.gov/ncbddd/autism/data.html

[2] W. Chung. How Big Data Are Unlocking the Mysteries of Autism, Scientific American Newsletter, Neurology Opinions. [Online]. Available: https://www.scientificamerican.com/article/how-big-dat a-are-unlocking-the-mysteries-of-autism/

[3] USF Health. Data Analytics Has Created a New Understanding of Autism. [Online]. Available: https://www.usfhealthonline.com/resources/healthcare-a nalytics/data-analytics-has-created-a-new-understandin g-of-autism/

[4] M. C. Christina, "Early intervention in autism, infants and young children," *Interdisciplinary Journal of Early Childhood Intervention*, vol. 18, pp. 74−85, 2005.

[5] A. Abdulrazzaq, S. Hamid, and A. A. Douri, "Early detection of autism spectrum disorders (asd) with the help of data mining tools," *Journal of BioMed Research International*, 2022.

[6] F. Thabtah, R. Spencer, and T. Dayara, "Autism screening: An unsupervised machine learning approach," *Health Information Science and Systems*, vol. 10, 2022.

[7] U. Madhuri. Predicting behavioral Challenges in ASD children. [Online]. Available: https://www.kaggle.com/datasets/uppulurimadhuri/datas

[8] L. Madhuri. ASD Children traits. [Online]. Available: https://www.kaggle.com/datasets/uppulurimadhuri/datas et

[9] J. Cai, J. Luo, S Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, 2018.

[10] M. Hall, "Correlation-based feature selection for machine learning, research commons," The University of Waikato, 2022

[11] S. Suresh, D. Newton, T. Everett, G. Lin, and B. Duerstock, "Feature selection techniques for a machine learning model to detect autonomic dysreflexia," *Front Neuroinform*, vol. 16, August 2022.

[12] J. Brownlee. How to Perform Feature Selection with Categorical. [Online]. Available: https://machinelearningmastery.com/feature-selection-w ith-categorical-data/

[13] P Sedgwick, *Pearson's Correlation Coefficient*, 2008.

[14] Random Forest Classifier. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn

[15] D. Moore, N. William, and M. Fligner, *The Basic Practice of Statistics*," W. H. Freeman Publisher, 2021.

[16] J. Brownlee, Information Gain and Mutual Information for Machine Learning. [Online]. Available: https://machinelearningmastery.com/information-gain-a nd-mutual-information/